



جامعة أبو بكر بلقايد - تلمسان

Université Abou Bakr Belkaïd de Tlemcen

Faculté de Technologie

Département de Génie Biomédical

Laboratoire de Recherche de Génie Biomédical

**MEMOIRE DE PROJET DE FIN D'ETUDES**

Pour l'obtention du Diplôme de

**MASTER en GENIE BIOMEDICAL**

*Spécialité* : Informatique Biomédicale

Présenté par : BOUDLAL Khadidja

---

**REALISATION D'UNE APPLICATION POUR LA  
DETECTION DE TYPE DE LA MASSE MAMMAIRE**

---

Soutenu le 25 Mai 2016 devant le Jury

M.	EL HABIB DAHO M.	<i>MAB</i>	Université de Tlemcen	Président
M.	MOUSSAOUI Djilali	<i>MAA</i>	Université de Tlemcen	Encadreur
Mme.	MEKKIOUI N. née HEDEILI	<i>MAA</i>	Université de Tlemcen	Examineur

Année universitaire 2015-2016

*" Une personne qui n'a jamais commis d'erreurs,  
N'a jamais tenté d'innover"  
Albert Einstein.*

*Je dédie ce modeste travail à  
Mes très chers parents qui m'ont éclairés mon chemin et qui m'ont encouragés et  
soutenu toute au long de mes études, J'espère que Dieu tout puissant me donne la  
force et le courage pour que je puisse rendre leurs sacrifices,  
Mes frères et mes sœurs,  
Tous mes amis,  
Mes enseignants,  
Mes collègues de promotion,  
A tous ceux qui me connaissent de près ou de loin,  
Et à tous ceux qui occupent une place dans mon cœur et plus particulièrement  
Chahinaz et Amani, Merci d'être toujours là pour moi.  
Qu'ils trouvent ici l'expression de toute ma reconnaissance.*

# Remerciements

*Je tiens tout d'abord à remercier الله القدير الرحيم qui m'a donné la force et la patience d'accomplir ce modeste travail.*

*En second lieu, je tiens à remercier mon encadreur Monsieur MOUSSAOUI DJILALI MMA à l'Université de Tlemcen, de ses précieux conseils et son aide durant toute la période du travail.*

*Je remercie également Monsieur BECHAR AMINE pour sa disponibilité, ses conseils et ses suggestions.*

*Je tiens à remercier aussi Monsieur EL HABIB DAHO M. pour l'intérêt porté à ce travail et d'avoir accepté de présider ce jury.*

*Ensuite je désire adresser mes sincères sentiments à Madame MEKKIOUI N. née HEDEILI d'avoir examiné ce travail.*

*Je remercie infiniment tous les enseignants du Département GBM qui m'a aidée durant tout mon cycle d'études.*

# Résumé

L'objectif de ce projet de fin d'étude est de réaliser un processus d'aide à la décision pour accompagner le praticien dans la détection de type d'une masse mammaire après la classification BIRADS, ce processus sera conçu par l'extraction de la connaissance à partir d'une base de données de patientes en mettant en application la classification.

Dans le cadre de notre travail nous nous intéresserons à la construction d'un arbre de décision optimal pour la classification de type d'une masse mammaire, en utilisant différents types d'algorithmes et en appliquant notre propre démarche, nous transformons ensuite cet arbre à un ensemble de règles afin de les intégrer dans une simple application pour être facile à utiliser.

## **Mots clés**

Cancer de sein, masse mammaire, système d'aide à la décision médicale, classification, arbre de décision, C4.5, ID3, RF (Forêt aléatoire), C-RT (CART), Rnd Tree, BIRADS.

# Abstract

The objective of this project is to realize a process of decision support to accompany the practitioner in the detection of type of a mammary mass after BIRADS classification; this process will be designed by the extraction of the Knowledge from patients' database by applying the classification.

Within the framework of our work we shall be interested in the construction of an optimal decision tree for type of a mammary mass classification, by using various types of algorithm and applying our own approach, we transform then this tree in a set of rules to integrate them into a simple application to be easy to use.

## Keywords

Breast cancer, mammary mass, decision support system, classification, decision tree, C4.5, ID3, RF (Random Forest), C-RT (CART), Rnd Tree, BIRADS.

## ملخص

الهدف من هذا المشروع هو توفير طريقة لدعم الطبيب المختص في اتخاذ القرار فيما يخص نوع الكتلة الموجودة في الثدي بعد استعمال برنامج BIRADS للتصنيف، ويتم تطوير هذه العملية باستخراج المعرفة انطلاقا من قاعدة بيانات خاصة بمرضى سرطان الثدي مستعملين التصنيف. في نطاق عملنا هذا سنركز على بناء شجرة القرارات المثلى لتصنيف نوع الكتلة الموجودة في الثدي، باستخدام أنواع مختلفة من الخوارزميات وتطبيق طرقنا الخاصة، بعد ذلك يتم تحويل هذه الشجرة إلى مجموعة من القواعد ليتم دمجها في تطبيق بسيط لتسهيل استعمالها.

## كلمات مفتاحيه

سرطان الثدي، كتلة، نظام مساعدة لاتخاذ القرارات الطبية، تصنيف، شجرة القرارات، C4.5 ، ID3 ، RF (غابة عشوائية)، Rnd Tree, C-RT، BIRADS.

# Table des matières

Remerciements .....	iii
Résumé.....	iv
Abstract .....	v
Table des matières.....	vi
Table des figures .....	viii
Liste des tableaux .....	ix
Liste des abréviations .....	x
Introduction générale .....	1
Chapitre 1 Contexte médicale et problématique.....	3
Partie 1 Cancer du sein.....	4
1 Le cancer.....	4
2 Anatomie du sein .....	6
3 Cancer du sein .....	7
Partie 2 La mammographie et le dépistage du cancer du sein .....	12
1 la Mammographie.....	12
2 Les pathologies mammaires .....	15
2.1 Les Microcalcifications Mcs.....	15
2.2 Les masses .....	15
2.2.1 La forme .....	16
2.2.2 Le contour (marge).....	16
2.2.3 La densité .....	17
3 La classification des pathologies mammaires .....	18
La classification BIRADS.....	18
Problématique et solution.....	19
Chapitre 2 Outils et Etat de l'art.....	21
Partie 1 Les SADM et la classification .....	22
1 les SADM .....	22
2 La classification.....	22
Partie 2 Choix de la base de données et des méthodes.....	25
1 Description de la base de données .....	25
2 L'état de l'art sur la base de données.....	27
3 Discussion et choix de la méthode .....	29
4 Les arbres de décisions .....	30
5 Construction d'un arbre de décision .....	32
5.1 L'algorithme ID3.....	33
5.2 L'algorithme C4.5 .....	34
5.3 L'algorithme C-RT .....	34
5.4 L'algorithme Rnd Tree .....	35
5.5 L'algorithme RF .....	35

Chapitre 3 Expérimentation et réalisation de l'application .....	37
1 Prétraitement de la base de données.....	38
1.1 Méthodes pour traiter les données manquantes .....	39
1.2 Etapes de prétraitement .....	39
1.3 Travail effectué.....	40
2 Extraction de l'arbre optimal .....	46
2.1 Classifieurs utilisés .....	46
2.2 Echantillonnage utilisé .....	46
2.3 Travail effectué.....	46
2.3.1 Classification par l'algorithme C4.5.....	46
2.3.2 Classification par l'algorithme ID3 .....	47
2.3.3 Classification par l'algorithme C-RT .....	47
2.3.4 Comparaison des Résultats .....	48
2.3.5 Classification par l'algorithme Rnd Tree .....	49
2.3.6 Classification par l'algorithme de RF.....	49
2.4 Résultat Final.....	51
3 Réalisation de l'application en implémentant l'arbre optimal .....	51
3.1 Description.....	52
3.2 Exécution .....	54
Conclusion générale.....	56
Annexes.....	I
Bibliographie.....	IX



# Table des figures

Figure 1.1 - Prolifération des cellules cancéreuses.....	5
Figure 1.2 - Anatomie du sein.....	6
Figure 1.3 - Localisation du cancer au niveau du sein.....	8
Figure 1.4 - Répartition du cancer de sein selon l'âge. ....	9
Figure 1.5 - Monarchie et cancer du sein.....	9
Figure 1.6 - Les composants d'un mammographe.....	12
Figure 1.7 - Exemples d'incidences en mammographie.....	14
Figure 1.8 - Les différentes formes possibles d'une masse. ....	16
Figure 1.9 - Les différents contours possibles d'une masse. ....	17
Figure 1.10 - Types de densité mammaire.....	18
Figure 2.1 - Histogramme représente le nombre des cas maligne ou bénigne en fonction de l'attribut âge. ....	26
Figure 2.2 - Exemple d'un arbre de décision. ....	31
Figure 2.3 - Un arbre de classification.....	31
Figure 2.4 - Un arbre de régression. ....	32
Figure 3.1 - Représentation de la procédure suivie dans le chapitre 3. ....	38
Figure 3.2 - Arbre construit par C4.5 pour prédire la valeur de l'attribut Densité.....	41
Figure 3.3 - Arbre construit par C4.5 pour prédire la valeur de l'attribut BIRADS. ...	42
Figure 3.4 - Description de l'arbre construit par C4.5 pour prédire la valeur de l'attribut Age. ....	43
Figure 3.5 - Description de l'arbre construit par C4.5 pour prédire la valeur de l'attribut Forme.....	43
Figure 3.6 - Description de l'arbre construit par C4.5 pour prédire la valeur de l'attribut Contour.....	44
Figure 3.7 - Interface de l'application PDM. ....	45
Figure 3.8 - Exécution de l'application PDM. ....	45
Figure 3.9 - Histogramme représente le TC obtenu par chaque algorithme.....	48
Figure 3.10 - Courbe représente le TC obtenu en fonction de nombre d'arbre par RF. ....	49
Figure 3.11 - Histogramme représente le temps de calcul par chaque algorithme pour donner le meilleur TC. ....	50
Figure 3.12 - Description de l'arbre optimal. ....	51
Figure 3.13 - Interface finale de l'application.....	52
Figure 3.14 - Icône de l'application DTMM.....	52
Figure 3.15 - La fenêtre Aide de l'application DTMM.....	53
Figure 3.16 - Exécution de l'application. ....	54

# Liste des tableaux

Tableau 1.1 - Atténuation radiologique des composants mammaires. ....	14
Tableau 1.2 - Conduite à tenir pour chaque classe de l'ACR selon la classification BIRADS.....	19
Tableau 2.1 - les définitions des grandeurs vp, vn, fp et fn. ....	25
Tableau 2.2 - Tableau Descriptive des intervalles de chaque attribut. ....	26
Tableau 2.3 - Information sur les exemples de la base MMD. ....	26
Tableau 2.4 - Nombre de données manquante pour chaque attribut.....	27
Tableau 2.5 - Résumé des résultats à partir de l'état de l'art sur la base de données MMD.....	28
Tableau 3.1 - TC obtenu par chaque algorithme pour l'attribut Densité.....	41
Tableau 3.2 - TC obtenu par chaque algorithme pour l'attribut BIRADS. ....	42
Tableau 3.3 - TC obtenu par chaque algorithme pour l'attribut Age. ....	42
Tableau 3.4 - TC obtenu par chaque algorithme pour l'attribut Forme. ....	43
Tableau 3.5 - TC obtenu par chaque algorithme pour l'attribut Contour.....	44
Tableau 3.6 - TC obtenu par l'algorithme C4.5 en utilisant différentes partition. ....	47
Tableau 3.7 - TC obtenu par l'algorithme ID3 en utilisant différentes partition. ....	47
Tableau 3.8 - TC obtenu par l'algorithme C-RT en utilisant différentes partition.....	47
Tableau 3.9 - TC obtenu par l'algorithme Rnd Tree en utilisant différentes partition. ....	49

# Liste des abréviations

ACR : American College of Radiology.

ANAES : Agence Nationale d'Accréditation et d'Evaluation en Santé.

ANN : le classifieur réseau de neurone.

BDD : Base De Données.

BIRADS : Breast Imaging Reporting and Data System.

CC : Cranio Caudale.

C-RT : Classification and Regression Tree.

DM : Données Manquantes.

DT : Decision Tree (arbre de décision).

DTMM : Détection de Type d'une Masse Mammaire.

GP : Programme Génétique.

GPD : Programme Génétique Distribué.

HNN : réseau de neurone hybride.

ID3 : Inductive Decision tree.

IEEE : Institute of Electrical and Electronics Engineers.

MMD : Mammographic Mass Data.

MLO : Médio Latérale Oblique.

PDM : Prédiction des Données Manquantes.

Rnd Tree : Random Tree (Arbre Aléatoire).

RF : Random Forest (Foret Aléatoire).

SADM : Système d'Aide à la Décision Médicale.

SADMD : Système d'Aide à la Décision Médicale pour le Diagnostic.

SVM : Séparateur à Vaste Marge.

TC : Taux de Classification.

UCI : University California Irvine.

# Introduction générale

Le cancer du sein est le cancer le plus courant chez la femme dans le monde, c'est l'un des principales causes de mortalité féminine.

La mammographie est la technique d'imagerie la plus sensible pour détecter ce type de cancer dans un stade précoce. Dans les pays développés, cette technique est complétée par un système de classification (BIRADS) qui permet de fournir un rapport représente le degré de suspicion concernant la masse mammaire et les directives et les recommandations associées.

Le système BIRADS peut fournir une de sept classes de degré de suspicion, donc notre but dans ce modeste travail de fin d'étude est d'avoir deux classes au lieu de sept, c'est-à-dire au lieu d'avoir une suspicion concernant une masse mammaire, on va essayer d'avoir une décision finale, alors que la masse mammaire soit maligne, soit bénin.

Plusieurs chercheurs ont travaillé sur ce sujet, l'idéal est d'étudier les travaux de ces chercheurs et de les améliorer. Pour atteindre notre but, on se base sur la fouille de données (Data Mining).

Notre plan de mémoire se compose de trois chapitres, le premier chapitre est une représentation du contexte médicale et la problématique posée, se divise en deux parties, la première partie est concernant le cancer du sein, et la deuxième représente la mammographie et le dépistage du cancer de sein en introduisant la classification BIRADS.

Le deuxième chapitre c'est pour le choix des outils en utilisant l'état de l'art concernant la problématique posée, se divise aussi en deux parties, la première partie est une représentation des SADM et de la classification, et la deuxième partie est une explication de la base de données et les méthodes à utilisées.

Le troisième chapitre représente l'expérimentation et la réalisation de l'application, il représente les expérimentations réalisées, les résultats obtenus avec leurs interprétations et une étude comparative avec d'autres résultats de certains travaux portant sur le même sujet, et enfin la réalisation de l'application finale.

En dernier lieu, une conclusion générale et les perspectives à venir dans ce travail.

# Chapitre 1

Contexte médicale et problématique

## **Introduction**

Le cancer du sein est le cancer le plus commun chez la femme dans les pays développés, lorsque ce type de cancer est détecté dans un stade très développé, il va être se propager vers d'autres partie du corps, c'est pour cette raison il faut le détecter dans un stade précoce.

Dans la première partie de ce chapitre, on va commencer par présenter la notion du cancer, l'anatomie du sein et enfin le cancer du sein, ensuite, dans la deuxième partie on va aborder les outils d'imagerie médicale permettant le dépistage et le diagnostic de ce type de cancer notamment la mammographie, on va étudier par la suite, d'une façon non exhaustive les caractéristiques des lésions mammaires.

L'accent sera mis finalement sur La classification des pathologies mammaires afin d'établir notre problématique dans ce modeste travail.

## **Partie 1 Cancer du sein**

### **1 Le cancer**

#### **1.1 Définition**

Le mot "cancer" désigne plus de 200 maladies, chacune d'entre elles porte un nom différent : cancer du poumon, cancer du sein, leucémie, etc. Tous les cancers sont différents les uns des autres, mais ils ont une chose en commun : ils s'attaquent aux cellules.

Le corps humain est composé de millions de cellule, il y en a de toutes sortes : cellules de la peau, du poumon, du sang, etc. Elles sont tellement petites qu'on ne peut les voir qu'au microscope, ces cellules ne vivent pas aussi longtemps que nous. Quand elles commencent à être trop vieilles pour bien fonctionner, elles se divisent en deux pour former de nouvelles cellules.

Un cancer commence quand une cellule cesse de faire son travail de façon normale, après un certain temps, cette cellule se divise pour se reproduire et on a deux cellules anormales. Ces nouvelles cellules vont se reproduire à leur tour et ainsi de suite, jusqu'au moment où des milliers de cellules anormales commenceront à prendre la place des cellules normales (Figure 1.1). C'est ce qu'on appelle un cancer [1].

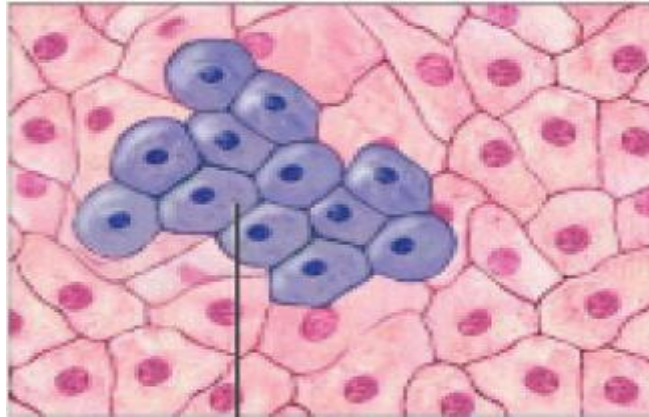


Figure 1.1 - Prolifération des cellules cancéreuses.

Alors, Une tumeur (cancer) est une masse qui se développe aux dépens d'un organe et à partir des cellules qui le constituent.

## 1.2 Types de cancer

Il existe deux types de cancer : bénigne et maligne.

- **Les tumeurs bénignes** ne sont pas cancéreuses, c'est à dire qu'elles n'envahissent pas les organes voisins et ne font que les repousser, elles ont un développement généralement limité, elles n'essaient pas leurs cellules ailleurs, ce qui signifie qu'elles ne font pas de métastases. Les tumeurs bénignes peuvent malgré tout poser des problèmes selon l'endroit où elles se situent (le tympan par exemple qu'elles peuvent détruire, ou les intestins qu'elles peuvent boucher).
- **Les tumeurs malignes** font exactement le contraire : elles envahissent toute la région, infiltrent les organes avoisinants et surtout elles envoient des métastases dans d'autres endroits du corps. Elles peuvent devenir énormes et récidivent souvent une fois qu'on les a retirées. Toutefois, ces tumeurs cancéreuses ne sont pas toutes mortelles, tout dépend de leur degré d'extension, de la précocité du traitement et du type de cellules qui les constituent [2].



## 2 Anatomie du sein

Les seins jouent un rôle important dans la féminité et dans l'image que la femme a de son corps, La fonction biologique du sein est de produire du lait afin de nourrir un nouveau-né.

La structure du sein est complexe, chaque sein (appelé aussi glande mammaire) est composé de quinze à vingt compartiments séparés par du tissu graisseux (adipeux) qui donne au sein la forme qu'on lui connaît, chacun de ces compartiments est constitué de lobules (tissu glandulaire) et de canaux lactifères (Figure 1.2), Le rôle des lobules est de produire le lait en période d'allaitement, les canaux transportent ensuite le lait vers le mamelon.

Les tissus mammaires sont influencés par des hormones produites par les femmes en quantité variable tout au long de leur vie (puberté, grossesse, allaitement...), Ces hormones sont l'œstrogène et la progestérone [3].

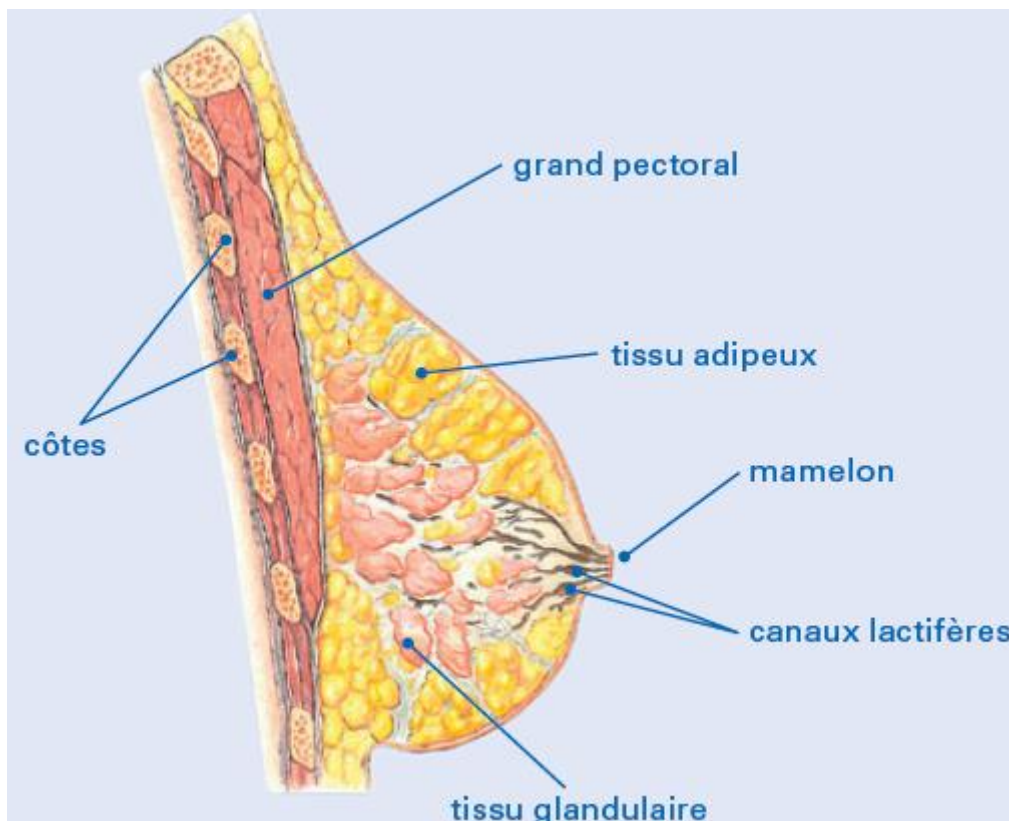


Figure 1.2 - Anatomie du sein.

### 3 Cancer du sein

#### 3.1 Définition

Le cancer du sein naît des cellules de l'appareil sécrétoire du sein constituées des lobules et des canaux lactifères.

Les cellules malignes se multiplient de manière désordonnée jusqu'à créer une tumeur qui s'attaque aux tissus, aux organes sains voisins et peut propager des cellules cancéreuses dans tout l'organisme : on dit alors que la tumeur métastase.

Le cancer du sein est le plus fréquent des cancers de la femme, Une femme sur 10 est actuellement touchée par le cancer du sein, et 1 % des cancers du sein est développé par un homme [4].

#### 3.2 Causes

- 5 à 10 % des cancers du sein diagnostiqués sont des cancers à prédispositions génétiques (*cancers familiaux*), Ce type de cancer apparaît par l'héritage.
- *Les cancers non-familiaux* constituent les 90 à 95 % restants, causé par plusieurs facteurs:
  - ✓ Causes hormonales (hyper-oestrogénémie) la stimulation oestrogénique (ce risque augmente en cas de puberté précoce (avant 12 ans), ménopause tardive (après 55 ans), ou première grossesse tardive).
  - ✓ Non-fécondité ou fécondité tardive: Les femmes qui n'ont pas eu d'enfant, ou qui ont eu leur première grossesse tardivement (après 30 ans) ont un risque sensiblement augmenté de développer un cancer du sein.
  - ✓ Polluants et autres perturbateurs endocriniens.
  - ✓ Obésité ou surpoids.
  - ✓ Absence d'allaitement.
  - ✓ Acides gras animaux, acides gras saturés (consommation des graisses animales).
  - ✓ Consommation d'alcool et de tabac.
  - ✓ Manque de vitamine D.

- ✓ Mastopathies: C'est un terme peu précis désignant toute maladie du sein, On le réserve en général à des anomalies bénignes qui peuvent prêter à confusion avec une tumeur.
- ✓ La modification du mode de vie (le stress, la sédentarité, ...).
- ✓ Alimentation (riche en graisse et pauvre en fibre).
- ✓ Les antibiotiques doublent le risque du cancer du sein: les femmes qui ont pris des ATB plus de 500 jours sur une période moyenne de 17ans ont 2 fois plus de risque de développer un cancer du sein.

### 3.3 Symptômes

Le cancer du sein se manifeste en général par la présence d'une boule dans le sein. Chez certaines patientes, il peut se signaler par un écoulement du mamelon, une présence de plaques rouges sur le sein, de crevasses, des plis anormaux ou d'une peau qui pèle, ... Une proportion importante de patientes ne présente pas de signes, mais uniquement des anomalies visibles sur une mammographie.

### 3.4 Statistiques

A partir d'un document a pour objectif principal de mesurer la fréquence du cancer du sein chez les femmes hospitalisées en gynécologie CHU Tlemcen pendant l'année 2006 [5] on a les résultats suivants:

- Localisation du cancer du sein : La coté droite représente 52% (Figure 1.3).

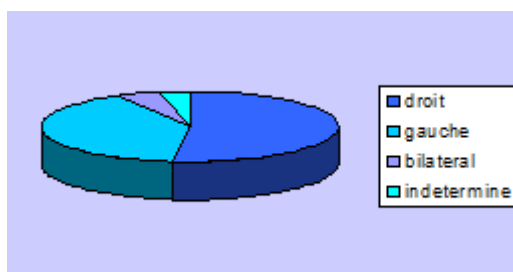


Figure 1.3 - Localisation du cancer au niveau du sein.

- Age et cancer du sein: L'âge le plus fréquent est de 40 a 50ans (Figure 1.4).

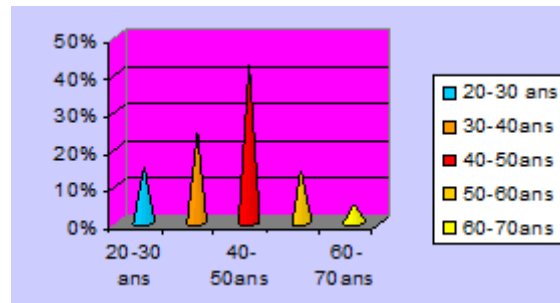


Figure 1.4 - Répartition du cancer de sein selon l'âge.

- La contraception et cancer du sein
  - ✓ 56.7% des femmes étaient sous contraception orale.
  - ✓ 40% utilisaient autres moyens.
  - ✓ 3,3% ne prenaient rien.
- Monarchie et cancer du sein (Figure 1-5): La très grande portion des femmes ayant une monarchie de 13ans.

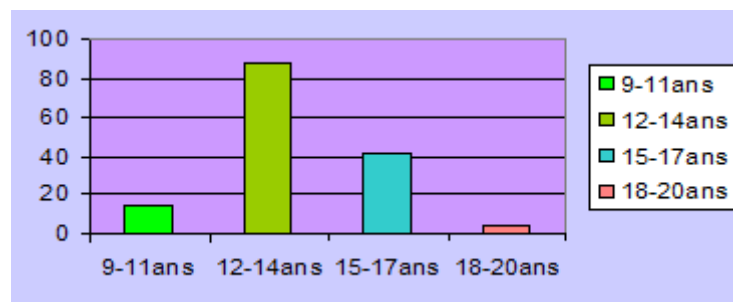


Figure 1.5 - Monarchie et cancer du sein.

## 3.5 Diagnostic

### 3.5.1 Les examens Principaux

Les examens nécessaires au diagnostic du cancer du sein sont de deux types :

- Les examens radiologiques, tels que la mammographie, les ultrasons (échographie), qui fournissent des images de l'intérieur du sein.
- Le prélèvement de tissus, tel que la biopsie, qui permet de confirmer ou d'exclure un cancer par une analyse au microscope des cellules prélevées.

Ces examens permettent d'établir le bilan diagnostique, qui a pour but de confirmer ou d'exclure que l'anomalie suspecte est un cancer.

S'il s'agit d'un cancer du sein, des examens complémentaires sont alors proposés pour déterminer le type de cancer et pour évaluer son étendue. Le choix de ces examens complémentaires dépend également de l'état de santé général de la patiente et du stade de la maladie.

### **3.5.1 Les examens complémentaires**

Lorsque la biopsie confirme que l'anomalie détectée dans le sein est bien un cancer, des examens complémentaires sont nécessaires, Les plus fréquents sont les suivants:

- Le bilan sanguin.
- L'imagerie par résonance magnétique (IRM).
- La scintigraphie.
- Le CT scan.
- Le PET-CT.

Ces examens complémentaires pour suivre trois buts principaux :

- ✓ Obtenir davantage de précisions concernant le type de tumeur (taille, emplacement, etc.) et son étendue.
- ✓ Déterminer si des cellules cancéreuses se sont déjà propagées dans d'autres parties du corps.
- ✓ Elaborer la stratégie thérapeutique, c'est-à-dire le traitement le plus approprié au type de cancer découvert et à son étendue, adapté à la situation de la patiente.

## **3.6 Traitement**

La grande majorité des cancers du sein peut aboutir à une guérison et les caractéristiques de la tumeur déterminent les choix du traitement.

### 3.6.1 La chirurgie

- ✓ Tumorectomie: chirurgie conservatrice, permet d'enlever une tumeur d'une taille habituellement inférieure à 3 cm et de conserver le sein.
- ✓ Mastectomie: retire le sein avec la tumeur, pour des tumeurs plus volumineuses ou s'il existe plusieurs tumeurs dans le sein.

### 3.6.2 Radiothérapie

La radiothérapie du sein permet de consolider l'effet de la chirurgie, Ce traitement s'applique sur le sein concerné, si celui-ci n'a pas été enlevé et permet de détruire les cellules cancéreuses grâce aux irradiations délivrées.

Il s'effectue en général sur une durée de 5 à 6 semaines et ne nécessite pas d'hospitalisation.

### 3.6.3 Chimiothérapie

La chimiothérapie qui permet la diffusion de médicaments destinés à détruire les cellules tumorales, est réalisée le plus souvent avant l'opération chirurgicale, Ce traitement n'est pas proposé lorsque par exemple la tumeur mesure moins d'un centimètre et que les ganglions sentinelles, situées proches de la tumeur ne sont pas atteints.

### 3.6.4 Hormonothérapie

Ce traitement qui consiste à délivrer des molécules qui bloquent les effets des œstrogènes sur la croissance des cellules cancéreuses, Ces produits sont proposés chez les femmes présentant un cancer hormono-dépendant du sein qui possède des récepteurs pour les œstrogènes [6].

- ✓ *On va détailler dans la partie suivante la mammographie pour établir la relation entre cette technique et la détection de type d'une mass mammaire.*

## Partie 2 La mammographie et le dépistage du cancer du sein

### 1 la Mammographie

La mammographie est une technique de radiographie, particulièrement adaptée aux seins de la femme, Elle a pour but de détecter au plus tôt des anomalies avant qu'elles ne provoquent des symptômes cliniques.

La mammographie est non seulement pratiquée dans les campagnes de dépistage du cancer du sein, mais aussi pour le diagnostic et la localisation lors d'interventions chirurgicales (ponctions), Le point fort d'un tel examen est qu'il permet d'examiner la totalité du tissu mammaire avec une ou deux incidences seulement.

L'appareil dédié à la réalisation d'une mammographie est le mammographe (Figure 1.6). Cet appareil se compose d'un tube radiogène générateur de rayons X de faible énergie (entre 20 et 50 kV) et d'un système de compression du sein.

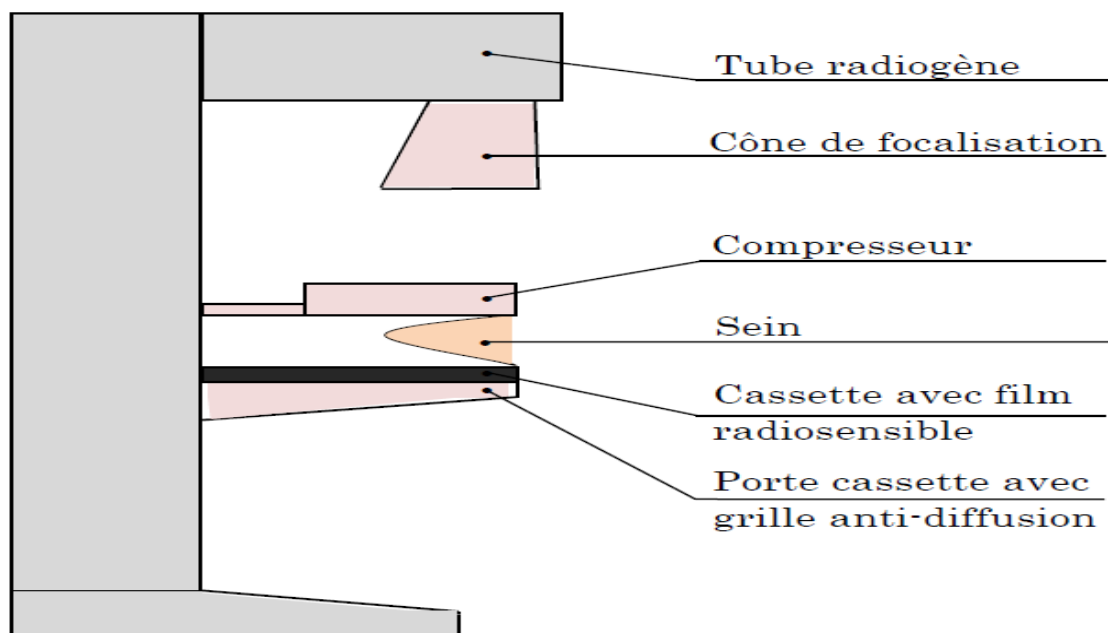


Figure 1.6 - Les composants d'un mammographe.

## 1.1 Déroulement de l'examen

En premier temps, les deux seins sont comprimés à tour de rôle, Cette compression permet l'étalement des tissus mammaires ce qui facilite la visualisation des structures du sein et la réduction de la dose de rayons X délivrée.

En deuxième temps, les deux seins sont exposés a une faible dose de rayons X, On obtient, alors, une projection du sein sur un détecteur plan. La radiographie est réalisée sur des films argentiques ou sur des systèmes de radiologie digitale de haute qualité.

L'analyse de la glande mammaire est réalisée grâce aux différences de l'atténuation des différents types de tissu.

## 1.2 Corrélation entre l'anatomie et les images mammographique

L'image mammographique est le résultat d'atténuation d'un faisceau de rayons X traversant les différents tissus mammaires, L'atténuation de ce faisceau dépend essentiellement de la composition des tissus traversés, En effet, la graisse est considérée comme une zone radio transparente vu qu'elle a une densité physique très légère, De ce fait, elle apparait très sombre sur un cliché mammographique, En revanche, les zones radio opaques apparaissent claires et correspondent au tissu fibroglandulaire et au calcium qui est le composant essentiel des lésions mammaires. Pour les matières prédominantes dans le sein, nous obtenons le Tableau 1.1 de correspondance entre les composants du tissu mammaire, la radio opacité et l'aspect sur le cliché mammographique.

En rassemblant les informations concernant l'anatomie et la radio transparence, on peut confirmer que l'aspect général d'une mammographie est sombre alors que les zones contenant des microcalcifications ou des masses (composées de calcium) sont plus claires.



Composant	Atténuation radiologique	Aspect sur mammographie
Graisse	radio transparent	très sombre
Eau	légèrement radio opaque	sombre
tissu conjonctif	radio opaque	Claire
calcium	très radio opaque	très claire

Tableau 1.1 - Atténuation radiologique des composants mammaires.

### 1.3 Les incidences en mammographie

Etant donné la complexité de l'anatomie du sein, la mammographie est généralement prise sous différentes directions appelées incidences, Une bonne incidence a pour but de visualiser le maximum de tissu mammaire en l'étalant le plus possible sur la plaque radiographique.

Selon la partie du sein à laquelle s'intéresse l'examen, différentes incidences sont utilisées, la Figure 1.7 explique le positionnement du tube radiogène et du détecteur pour les différentes incidences.

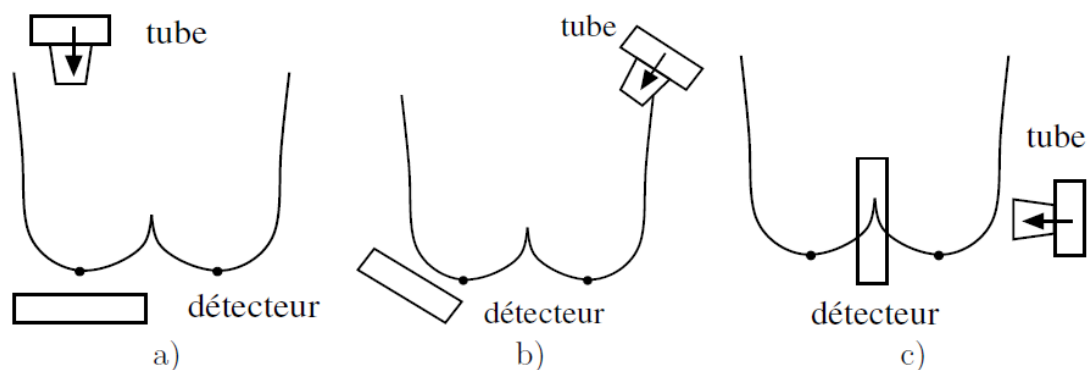


Figure 1.7 - Exemples d'incidences en mammographie.

a) face CC, b) MLO et c) Profil interne.

### 1.4 Les examens mammographiques

Vu son importance, la mammographie est actuellement pratiquée dans deux circonstances : dans le cadre d'un dépistage ou d'un diagnostic. Elle est aussi pratiquée lors d'une biopsie (prélèvement d'un petit morceau du tissu de l'anomalie et

son analyse au microscope) ou bien pour la localisation d'une lésion lors d'une intervention chirurgicale.

### **1.4.1 Le dépistage**

Le dépistage du cancer du sein consiste à pratiquer des examens de contrôle qui permettent de mettre en évidence des anomalies sans même la présence de symptômes décelables.

### **1.4.2 Le diagnostic**

La mammographie diagnostique est généralement réalisée après un examen de dépistage. L'objectif principal de cette mammographie de diagnostic est soit la recherche d'un signe radiologique dans une zone suspecte, soit l'analyse d'une façon plus précise d'une lésion détectée cliniquement (douleur, écoulement du mamelon, rougeur ou rétrécissement de la peau, palpation d'une lésion...).

## **2 Les pathologies mammaires**

### **2.1 Les Microcalcifications Mcs**

Une microcalcification est un dépôt de sels de calcium composé des substances chimiques  $\text{Ca}_3(\text{PO}_4)_2$ ,  $\text{CaCO}_3$  et  $\text{Mg}_3(\text{PO}_4)_2$ . Ces substances sont très radio-opaques et se traduisent, dans les clichés mammographiques, par de petits points clairs. Les caractéristiques qui distinguent les microcalcifications des autres éléments sont leur fort contraste et leur petite taille ( $< 0,5\text{mm}$ ). Une fois leur taille dépasse  $1\text{mm}$ , on les appelle des macrocalcifications et elles sont souvent bénignes.

### **2.2 Les masses**

Une opacité ou une masse est une lésion importante occupant un espace et vue sur deux incidences différentes. Si une opacité potentielle est vue seulement sur une seule incidence alors elle est appelée asymétrie jusqu'à ce que son caractère tridimensionnel

soit confirmé. Différentes caractéristiques de ces masses sont à décrire à savoir la forme, le contour et la densité.

### 2.2.1 La forme

Les masses mammaires peuvent avoir la forme ronde (Figure 1.8 a), ovale (figure 1.8 b), lobulée (figure 1.8 c) ou irrégulière (figure 1.8 d).

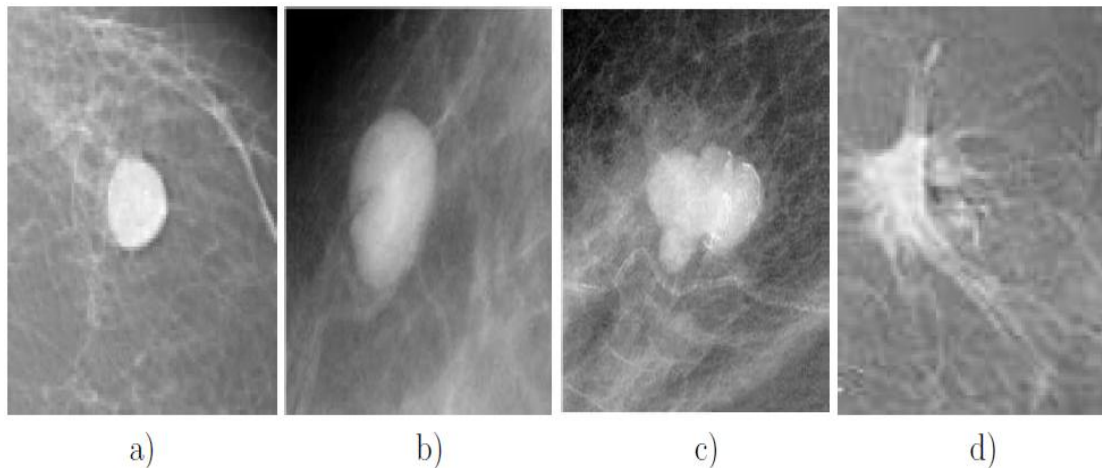


Figure 1.8 - Les différentes formes possibles d'une masse.

- a) **Ronde** : Il s'agit de masse sphérique, circulaire ou globuleuse.
- b) **Ovale** : Elle présente une forme elliptique (ou en forme d'œuf).
- c) **Lobulée** : La forme de la masse présente une légère ondulation.
- d) **Irrégulière** : Cette appellation est réservée aux masses dont la forme est aléatoire et ne peut être caractérisée par les termes cités ci-dessus.

### 2.2.2 Le contour (marge)

Le contour des masses mammaires est soit circonscrit (figure 1.9 a), soit microlobulé (figure 1.9 b), soit masqué (figure 1.9 c) soit indistinct (figure 1.9 d), soit spiculé (figure 1.9 e).

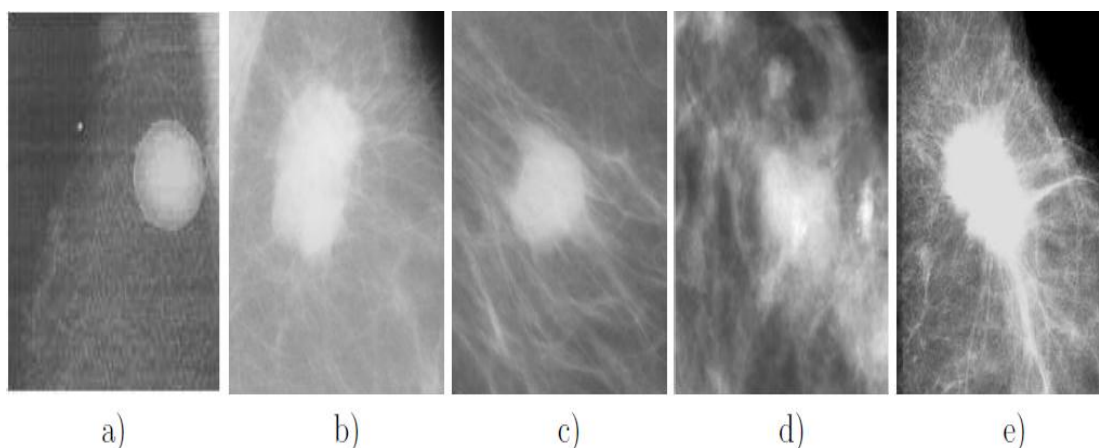


Figure 1.9 - Les différents contours possibles d'une masse.

- a) **Circinscrit** : Il s'agit d'une transition brusque entre la lésion et le tissu environnant, Le contour est alors net et bien défini. Pour qu'une masse soit qualifiée de circonscrite, il faut qu'au moins 75% de son contour soit nettement délimité.
- b) **Microlobulé** : Dans ce cas, de courtes dentelures du contour créent de petites ondulations.
- c) **Masqué** : Un contour masqué est un contour qui est caché par le tissu normal adjacent, Ce terme est employé pour caractériser une masse circonscrite dont une partie du contour est cachée.
- d) **Indistinct** : Dans ce cas, le contour est mal défini, Ce caractère indistinct (le contraire de circonscrit) peut correspondre à une infiltration.
- e) **Spiculé** : La masse est caractérisée par des lignes radiaires prenant naissance sur le contour de la masse, Ces lignes radiaires sont appelées les spicules.

### 2.2.3 La densité

L'aspect du sein normal est très variable d'une femme à l'autre, Le facteur le plus remarquable est la grande variabilité de la densité radiologique de l'aire mammaire. Il existe 4 classes de la composition du sein : Contient grasses (Figure 1.10 a), Basse (Figure 1.10 b), Moyen (Figure 1.10 c), Elevée (Figure 1.10 d).

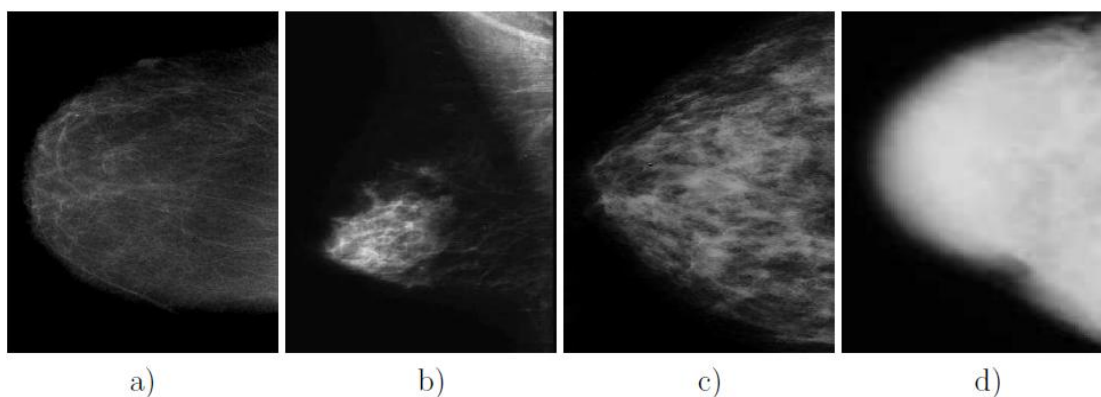


Figure 1.10 - Types de densité mammaire.

- a) **Contient graisses** : Le sein est presque entièrement graisseux et homogène, radio transparent et facile à lire (moins de 25 % de la glande mammaire).
- b) **Basse** : Il y a des opacités fibroglandulaires dispersées, Le sein est graisseux et hétérogène (approximativement 25 à 50 % de la glande mammaire).
- c) **Moyen** : Le tissu mammaire est dense et hétérogène (approximativement 51 à 75 % de la glande mammaire).
- d) **Elevée** : Le tissu mammaire est extrêmement dense et homogène. La mammographie est alors difficile à interpréter puisque la densité peut masquer une lésion (plus de 75 % de la glande mammaire).

### 3 La classification des pathologies mammaires

Il est important d'adopter un lexique standard et une classification commune afin de fournir aux radiologues une description claire et précise des lésions mammaires. L'étude morphologique de ces lésions a fait l'objet de plusieurs classifications à savoir la classification de Le Gal, de Lanyi et de BIRADS. Cette dernière est la plus connue et la plus pratiquée.

#### La classification BIRADS

L'ACR a souligné l'importance d'un protocole mammographique standardisé et complet qui tient en compte les différents facteurs de malignité. En novembre 1998, l'ACR a établi le système Américain BIRADS qui a été rédigé par un groupe d'experts réunis par l'ANAES.

Ce système permet de classer les images mammographiques en 7 catégories en fonction du degré de suspicion de leur caractère pathologique.

La classification de l'ACR résume les formes des différentes masses et des différentes microcalcifications, leur texture, les différents aspects de la distorsion architecturale ainsi que leur degré de malignité. Une fois que le radiologue arrive à reconnaître la catégorie d'une mammographie, il sait automatiquement les directives et les recommandations associées à cette classe (Tableau 1.2).

Classe BIRADS	Recommandation associée
ACR 0	<b><i>Incomplète : Nécessité d'investigations complémentaires</i></b> telles que cliché avec compression centrée, agrandissement, incidence particulière, échographie.
ACR 1	<b><i>Négative</i></b> : mammographie normale, aucune masse ou calcification suspecte n'est présente.
ACR 2	<b><i>Lésion bénigne</i></b> : ne nécessitant ni surveillance ni examen complémentaire
ACR 3	<b><i>Lésion probablement bénigne</i></b> : une surveillance clinique et radiologique à court terme est conseillée.
ACR 4	<b><i>Anomalie suspecte</i></b> : une biopsie devrait être envisagée. Ces lésions n'ont pas un aspect typique de cancer, mais peuvent néanmoins correspondre à une lésion maligne.
ACR 5	<b><i>Probablement maligne</i></b> : Lésions fort suspectes de malignité, l'anomalie est évocatrice d'un cancer.
ACR 6	<b><i>Lésion Maligne</i></b>

Tableau 1.2 - Conduite à tenir pour chaque classe de l'ACR selon la classification BIRADS.

Enfin la pratique a démontré que l'utilisation de ce système permet d'augmenter le taux de reconnaissance des masses malignes et celles bénignes [7].

## Problématique et solution

Près de 70% des biopsies demandées suite aux mammographies s'avèrent inutiles [8], alors le but de ce projet est de réaliser un processus d'aide à la décision pour accompagner le praticien dans le diagnostic du type d'une masse mammaire sans biopsie.

D'après la deuxième partie de ce chapitre, la classification de BIRADS permet d'augmenter le taux de reconnaissance des masses malignes et celles bénignes, ce qui nous permet de reconnaître le type de la masse mammaire dans un stade précoce.

Dans la suite de ce projet on va essayer d'améliorer ce taux de reconnaissance, en utilisant la classification de BIRADS et d'autres paramètres liés à une masse mammaire.

## **Conclusion**

Dans ce chapitre, la notion du cancer du sein ainsi que la mammographie ont été présentés. Une grande attention a été consacrée à l'étude des spécifications des pathologies mammaires à savoir les masses et les microcalcifications.

Une telle étude est fortement associée à la présentation des standards adoptés par les radiologues pour classifier les lésions mammaires en bénignes ou malignes. L'objectif de détailler ces études est de mieux introduire le choix de notre méthode d'amélioration pour détecter le type d'une masse mammaire.

Dans le chapitre suivant on va choisir et étudier les outils et les méthodes de classification nécessaires qui s'adaptent avec notre problématique.

# Chapitre 2

## Outils et Etat de l'art



## **Introduction**

Ce chapitre comporte deux parties : dans un premier temps, on va présenter les systèmes d'aide à la décision médicale et la classification pour mettre en évidence la relation entre ces deux concepts, La deuxième partie présente les outils telque la base de données adaptable à notre problématique et l'étude de l'état de l'art sur cette base de données afin de choisir les méthodes de classification à utiliser.

## **Partie 1 Les SADM et la classification**

### **1 les SADM**

Les SADM sont définis de manière très générale comme des outils informatiques dont le but est de fournir aux cliniciens en temps et lieux utiles les informations décrivant la situation clinique d'un patient ainsi que les connaissances appropriées à cette situation, correctement filtrées et présentées afin d'améliorer la qualité des soins et la santé des patients.

Il existe ainsi des SADM pour l'ensemble des activités médicales (prévention, dépistage, diagnostic, traitement) et la majorité des spécialités médicales (maladies chroniques ou affections aiguës) [9].

Ce que nous intéresse dans notre travail est les SADMD, Pour obtenir le résultat final à l'aide d'un SADMD il faut fournir des données caractérisant l'anomalie a détecté, afin de les utiliser pour classer le type de cette anomalie à l'aide d'un algorithme de classification.

### **2 La classification**

L'objectif de la classification est d'identifier les classes auxquelles appartiennent des objets à partir de traits descriptifs (attributs, caractéristiques), en utilisant un algorithme qui s'appel classifieur [10].

## 2.1 Les phases de classification

Le classifieur qui réalisera le classement doit passer par deux phases, une phase d'apprentissage et une phase de test.

### 2.2.1 Phase d'apprentissage

Le but de l'apprentissage est de découvrir les règles qui gouvernent et régissent des formes, L'apprentissage est un processus calculatoire qui doit être capable d'amener à une certaine prédiction et à une certaine généralisation.

Il existe trois types d'apprentissages principaux:

#### a. Apprentissage supervisé

C'est actuellement le mode d'apprentissage le plus couramment utilisé, Son principe est élémentaire : on soumet au classifieur un grand nombre d'exemples pour lesquels l'entrée et la sortie associée sont connues et les paramètres d'apprentissage sont modifiés de façon à corriger l'erreur commise par le classifieur (c'est-à-dire la différence entre la sortie désirée et la réponse du classifieur à l'entrée correspondante).

Le classifieur a alors comme but d'approximer ces exemples aussi bien que possible et de développer à la fois la bonne représentation mathématique qui lui permet de généraliser ces exemples pour ensuite traiter des nouvelles situations (qui n'étaient pas présentes dans les exemples).

#### b. Apprentissage non-supervisé

Contrairement aux modes supervisés, seule une base d'entrées est ici fournie au classifieur, Celui-ci doit donc déterminer lui-même ses sorties en fonction des similarités détectées entre les différentes entrées, c'est-à-dire en fonction d'une règle d'auto-organisation, Il n'y a donc pas de professeur, puisque c'est au classifieur de découvrir les ressemblances entre les éléments de la base de données.

#### c. Apprentissage par renforcement

L'apprentissage par renforcement consiste à apprendre quoi faire, comment associer des actions à des situations, afin de maximiser quantitativement une

récompense, On ne dit pas à l'apprenant quelle action faire, mais au lieu de cela, il doit découvrir quelles actions donnent le plus de récompenses en les essayant.

### 2.2.2 Phase de test

Cette phase doit permettre l'affectation d'un nouvel objet à l'une des classes, au moyen d'une règle de décision intégrant les résultats de la phase d'apprentissage, L'objectif est d'obtenir une estimation la plus fidèle possible du comportement du classifieur dans des conditions réelles d'utilisation, Pour cela, des critères classiques comme le TC et le taux d'erreur sont presque systématiquement utilisés, Mais d'autres critères, comme la spécificité et la sensibilité, apportent aussi des informations utiles.

#### a. Taux de classification

Le TC (1) et taux d'erreur permettent d'évaluer la qualité du classifieur par rapport au problème pour lequel il a été conçu, Ces taux sont évalués grâce à une base de test qui contient des formes étiquetées par leur classe réelle d'appartenance comme celles utilisées pour l'apprentissage afin de pouvoir vérifier les réponses du classifieur.

$$TC = \frac{vp(i) + vn(i)}{vp(i) + vn(i) + fp(i) + fn(i)} \dots\dots\dots(1)$$

#### b. Sensibilité et Spécificité

L'évaluation des performances d'un classifieur aussi peut être réalisée par l'appréciation de deux lois statistiques, qui sont la sensibilité (2) et la spécificité (3).

La sensibilité  $Se(i)$  représente la probabilité de bonne classification de la classe  $i$  et la spécificité  $Sp(i)$  est une mesure indirecte de la probabilité de fausse alarme égale à  $1 - Sp(i)$  [11].

$$Se(i) = \frac{vp(i)}{vp(i) + fn(i)} \dots\dots\dots(2)$$

$$Sp(i) = \frac{vn(i)}{vn(i) + fp(i)} \dots\dots\dots(3)$$

Avec les grandeurs  $vp(i)$ ,  $fn(i)$ ,  $vn(i)$ ,  $fp(i)$  sont définies dans le Tableau 2.1 :

	Présence d'événement de classe $i$	Absence d'événement de classe $i$
Classification Positive	Vrai Positif $vp(i)$	Faux Positif $fp(i)$
Classification Négative	Faux Négatif $fn(i)$	Vrai Négatif $vn(i)$

Tableau 2.1 - les définitions des grandeurs  $vp$ ,  $vn$ ,  $fp$  et  $fn$ .

### 2.3 Etapes d'apprentissage pour atteindre une bonne classification

1. Choisir une base de données qui est adaptable à la problématique posée.
2. Faire des prétraitements sur cette base de données (s'il y a des données manquantes par exemple, si la base de données contient des attributs inutiles, etc).
3. Appliquer une méthode de classification (classifieur) sur cette base de données, en utilisant des méthodes d'échantillonnage (par exemple la validation croisée), afin de vérifier les critères d'évaluation (TC, Se et Sp).
4. Selon les critères d'évaluation, construire une méthode (ou ensemble des règles) de classification automatique afin de l'utiliser pour classer de nouveaux cas.

## Partie 2 Choix de la base de données et des méthodes

Dans les différents travaux réalisés sur l'amélioration de la classification BIRADS pour détecter le type de la masse mammaire, on retrouve que les chercheurs travaillent sur la base de données MMD d'UCI.

### 1 Description de la base de données

- Le titre: Mammographic Mass Data (MMD) [12].
- Faite par Prof. Dr. Rüdiger Schulz-Wendtland, à l'institut de Radiologie, (Radiologie gynécologique), Université de Erlangen-Nuremberg-Allemand entre 2003 et 2006, elle est distribuée en Octobre 2007.
- La base de donnée contient 961 individus, sur les quelles est mesuré 5 attributs avec une classe de sortie (Tableau 2.2).

	0	1	2	3	4	5	6
BIRADS	Incomplète	Négative	Bénigne	Probablement bénigne	Suspecte	Probablement maligne	Maligne
L'Age	-	-	-	-	-	-	-
Forme	-	Ronde	Ovale	Lobulée	Irrégulière	-	-
Contour	-	Circonscrit	Micro-lobulé	Masqué	Indistinct	Spiculé	-
Densité	-	Elevée	Moyen	Basse	Contient graisses	-	-

Tableau 2.2 - Tableau Descriptive des intervalles de chaque attribut.

- Concernant l'intervalle de variance d'âge dans la base de données on a l'histogramme suivant (figure 2.1) [13].

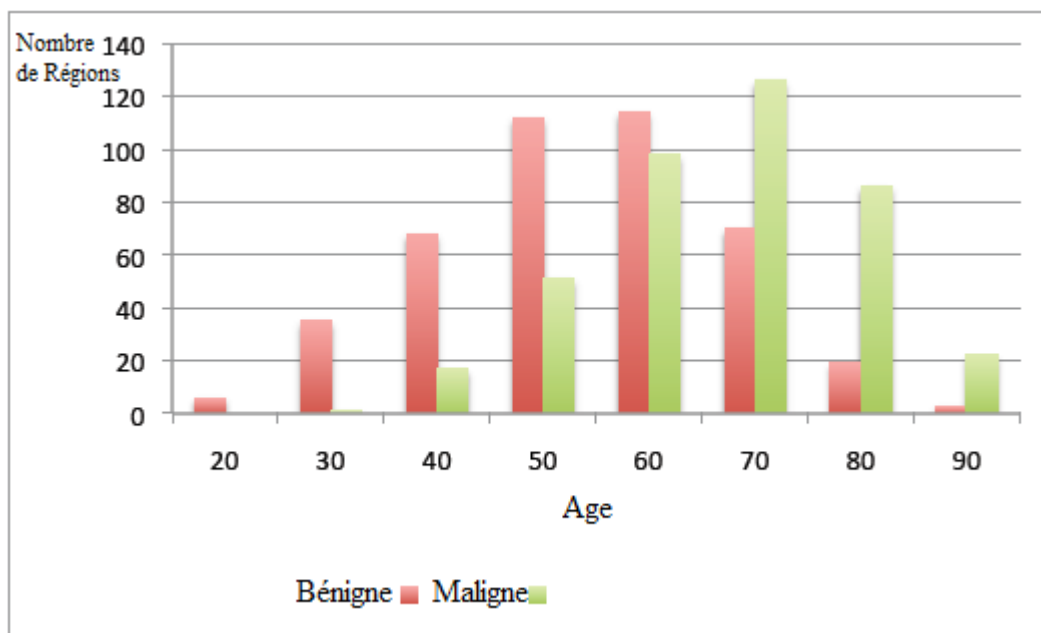


Figure 2.1 - Histogramme représente le nombre des cas maligne ou bénigne en fonction de l'attribut âge.

- Les informations concernant la classe de sortie sont représentées dans le tableau 2.3.

N° de la Classe	Label	Nombre
0	bénigne	516
1	maligne	445

Tableau 2.3 - Information sur les exemples de la base MMD.

- Dans le cadre d'étude de la base de données on a trouvé qu'elle contient des données manquantes (Tableau 2.4).

Attribut	BIRADS	Age	Forme	Contour	Densité
Nombre de données manquantes	2	5	31	48	76

Tableau 2.4 - Nombre de données manquante pour chaque attribut.

Dans le chapitre suivant on va faire des traitements sur la base de données pour résoudre ce problème des données manquantes.

- **Remarque**

Dans la base de données MMD, chaque observation (exemple) est associée à un label (sortie), pour cette raison on peut dire que la classification basée sur l'apprentissage supervisé est la classification qui s'adapte à notre travail.

## 2 L'état de l'art sur la base de données

1. Ali KELES et Ayturk KELES, en 2013 [14], ont travaillé sur la base de données MMD, ils ont appliqué la méthode "neuro-fuzzy", les résultats obtenus sont TC=84%, Sp=80% et Se=90%.

2. Shu-Ting Luo et Bor-Wen Cheng, en 2010 [15], ont travaillé sur la base de données MMD, ils ont appliqué les deux méthodes de classification DT et SVM les résultats obtenus sont TC = 86.6%, SP=82.9%, Se = 83.6% pour DT et TC = 81.3%, SP= 81% Se = 81.7% pour l'SVM.

3. Alaa M. Elsayad, en 2010 [16], a travaillé sur la base de données MMD, il a appliqué un ensemble de classifieur Bayésien, les résultats obtenus sont TC=90.63%, SP=94.12 % et Se=87.50%.

4. Benaki Lairenjam et Yengkhom Satyendra Singh, en 2015 [17], ont travaillé sur la base de données MMD, ils ont appliqué les trois méthodes de classification ANN,

Naive Bayes et HNN les résultats obtenus sont TC = 84% par ANN, TC = 81.874% par Naive Bayes et TC=87.42% par HNN.

5. Benaki Lairenjam et Siri Krishan Wasan, en 2009 [18], ont travaillé sur la base de données MMD, ils ont appliqué une classification basée sur l'association multiple des règle (CMAR), le résultat trouvé est un TC = 84.52%.

6. Kathleen H. Miao, and George J. Miao, Senior Member (IEEE Titre), en 2013 [19], ont travaillé sur la base de données MMD, ils ont appliqué la classification par ANN, enfin ils ont trouvé TC = 89.64%, Sp = 89.93%, Se = 89.33%.

7. Simone A. Ludwig, en 2010 [20], a travaillé sur la base de données MMD, il applique un GP et un GPD, il obtient comme résultats (d'après la courbe de Roc)  $A(z) = 85.9\%$ ,  $Spec_{0.95}=50.03\%$  et  $A(z)_{0.9} = 29\%$  par GP et  $A(z) = 86\%$ ,  $Spec_{0.95}=90\%$  et  $A(z)_{0.9} = 83.87\%$  par GPD.

• **On résume les résultats de l'état de l'art dans un tableau (Tableau 2.5)**

Etat de l'art	Méthode	TC	Sp	Se
1	neuro-fuzzy	84%	80%	90%
2	DT	86.6%	82.9%	83.6%
	SVM	81.3%	81%	81.7%
3	classifieur Bayésien	90.63%	94.12 %	87.50%
4	ANN	84%		
	Naive Bayes	81.874%	–	–
	HNN	87.42%		
5	CMAR	84.52%	–	–
6	ANN	89.64%	89.93%	89.33%
		A(z)	Spec <sub>0.95</sub>	A(z) <sub>0.9</sub>
7	GP	85.9%	50.03%	29%
	GPD	86%	51.4%	27.1%.

Tableau 2.5 - Résumé des résultats à partir de l'état de l'art sur la base de données MMD.

## 3 Discussion et choix de la méthode

### 3.1 Discussion sur les résultats

Selon le tableau au-dessus il est remarquable que les résultats et plus particulièrement le TC est différent d'un travail à un autre:

- Le plus petit TC (81.3%) est obtenu par le classifieur SVM dans le travail Shu-Ting Luo et Bor-Wen Cheng, en 2010.
- Le meilleur TC (90.63%) est obtenu par le classifieur Bayésien dans le travail d'Alaa M. Elsayad, en 2010.
- D'autres classifieurs sont utilisés sur la base de données MMD telque DT, neuro-fuzzy, ANN, Naive Bayes, HNN, CMAR, GP et GPD, les résultats de TC sont aussi acceptables (entre 81.874% et 89.64%).

### 3.2 Choix de la méthode

Il existe plusieurs manières pour choisir la méthode de travail, la manière la plus connue est de travailler par la méthode qui donne les meilleurs résultats selon l'état de l'art, et d'essayer d'améliorer ces résultats.

La méthode qui donne les meilleurs résultats et le classifieur Bayésien, mais notre problème avec ce classifieur est qu'il donne un modèle complexe qui n'est pas simple à l'interpréter.

Alors, dans ce modeste travail, on va travailler par les arbres de décision, à la suite on va essayer d'augmenter le TC de façon tel qu'il sera le plus grand possible que les résultats trouvés dans le travail de Alaa M. Elsayad. Notre choix est selon trois critères:

- ✓ Selon l'état de l'art, DT donne à Shu-Ting Luo et Bor-Wen Cheng un TC intéressant (86.6%).
- ✓ Les arbres de décisions donnent des résultats interprétables et faciles à les implémenter (dans différent langage telque C++, Un arbre de décision peut être facilement transformé en un ensemble de règles) afin de compléter le travail par une application simple et utilisable.
- ✓ La rapidité des DT (en terme de calcul et de donner les résultats).



## 4 Les arbres de décisions

### 4.1 Définition

Un arbre de décision est, comme son nom le suggère, un outil d'aide à la décision qui permet de répartir une population d'individus en groupes homogènes selon des attributs discriminants en fonction d'un objectif fixé et connu. Il permet d'émettre des prédictions à partir des données connues sur le problème par réduction, niveau par niveau, du domaine des solutions [21], d'autre façon c'est l'ensemble de règles de classification basant leur décision sur des tests associés aux attributs, organisés de manière arborescente [22].

L'arbre de décision est une méthode qui a l'avantage d'être lisible pour les analystes et permet de déterminer les couples "attribut, valeur" discriminants à partir d'un très grand nombre d'attributs et de valeurs.

### 4.2 Vocabulaire des arbres

- Un arbre est constitué de nœuds connectés entre eux par des branches.
- Une branche entre deux nœuds est orientée : l'un des nœuds de la connexion est dit nœud parent, et l'autre nœud enfant.
- Chaque nœud est connecté à un et un seul nœud parent, sauf le nœud racine qui n'a pas de parent.
- Chaque nœud peut être connecté à 0 ou n nœuds enfants.
- Les deux caractéristiques précédentes font qu'un arbre n'est pas un réseau ou graphe.
- Un nœud qui n'a pas de parents est appelé nœud racine ou racine.
- Un nœud qui n'a pas de nœuds enfants est appelé nœud feuille ou feuille [23] (Figure 2.2).

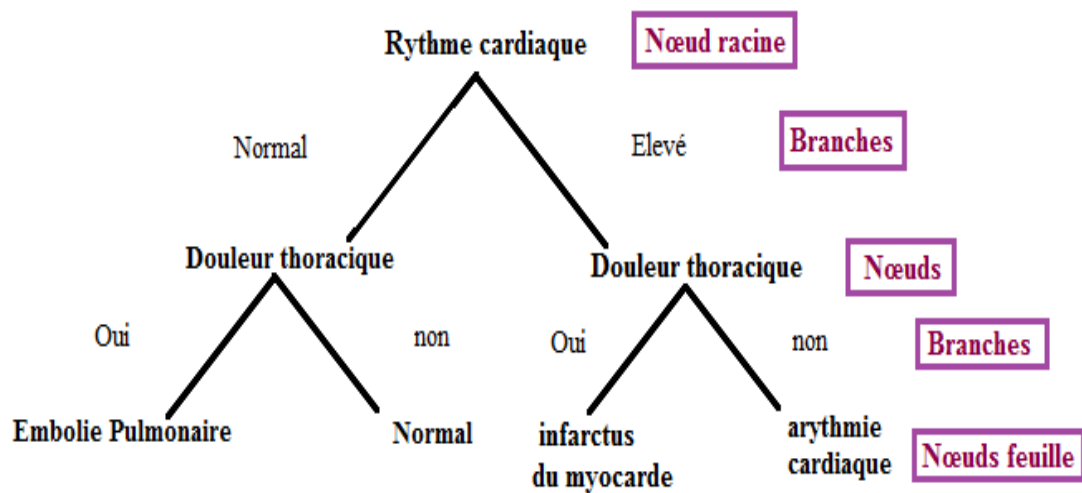


Figure 2.2 - Exemple d'un arbre de décision.

Chaque nœud interne d'un arbre de décision porte sur un attribut discriminant des éléments à classer qui permet de répartir ces éléments de façon homogène entre les différents fils de ce nœud, Les branches liant un nœud à ses fils représentent les valeurs discriminantes de l'attribut du nœud, Et enfin, les feuilles d'un arbre de décision sont ses prédictions concernant les données à classer [22].

### 4.3 Types d'arbres

Il existe deux types d'arbre de décision, arbre de classification et arbre de régression.

- Les arbres de classification permettent de prédire à quelle classe la variable cible appartient, dans ce cas la prédiction est une étiquette de classe (Figure 2.3).

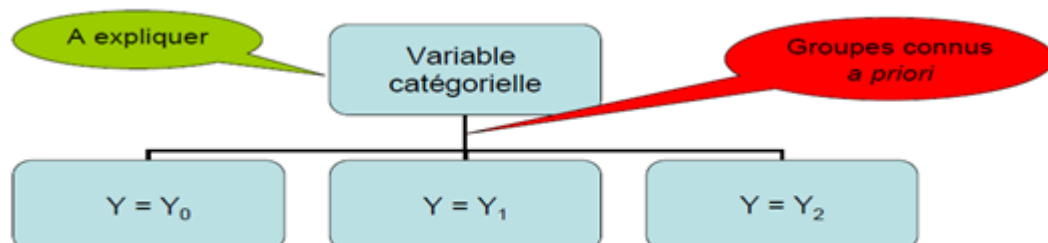


Figure 2.3 - Un arbre de classification.

Dans les arbres de classification les groupes définis a priori par la variable catégorielle à expliquer, C'est-à-dire quelles variables classent le mieux les observations dans les groupes? [24].

- Les arbres de régression permettent de prédire une quantité réelle (par exemple, la durée de séjour d'un patient dans un hôpital), dans ce cas la prédiction est une valeur numérique [25] (Figure 2.4).

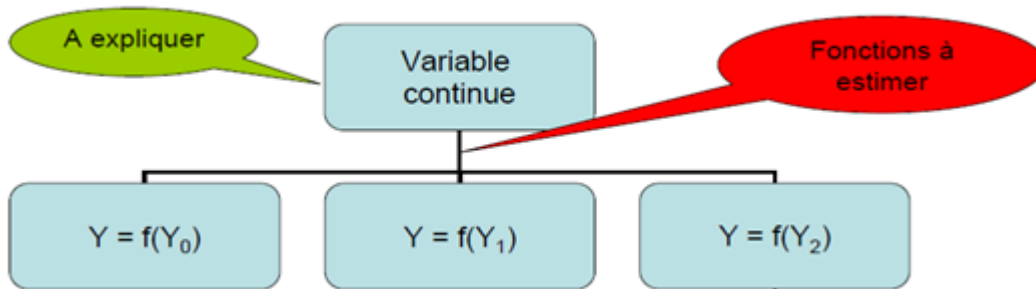


Figure 2.4 - Un arbre de régression.

Dans les arbres de régression les groupes définis à posteriori pour la variable continue à expliquer c'est-à-dire quelles variables prédisent le mieux la variable à expliquer? [24].

## 5 Construction d'un arbre de décision

Notre objectif c'est d'apprendre un arbre de décision à partir d'une base d'exemples étiquetés pour que cet arbre doive être efficace en généralisation (être capable de classer correctement un nouvel exemple).

Pour cette raison on mit l'accent sur la construction intelligente d'un arbre de décision (Algorithme) [26].

---

### **Algorithme** arbre de décision

---

On considère un nœud

On sélectionne un attribut pour ce nœud

On crée une branche pour chaque valeur de cet attribut

Pour chaque branche, on regarde la pureté de la classe obtenue

On décide si on termine la branche ou non

Si on ne termine pas le processus est répété

---

On dénombre plusieurs algorithmes pour construire des arbres de décision, parmi lesquels : ID3, C4.5, C-RT, Rnd Tree, et RF [27], Ces algorithmes sont très utilisés pour leur performance et par le fait qu'ils génèrent des procédures de classification exprimables sous forme de règles.

A la suite on va détailler d'une manière générale tous ces algorithmes.

## 5.1 L'algorithme ID3

L'algorithme ID3 a été développé à l'origine par Ross Quinlan, Il a tout d'abord été publié dans le livre Machine Learning en 1986 [28].

ID3 construit un arbre de décision de façon récursive en choisissant l'attribut qui a gain d'information (5) le plus élevé selon l'entropie de Shannon (4), Cet algorithme fonctionne exclusivement avec des attributs catégoriques et un nœud est créé pour chaque valeur des attributs sélectionnés.

### Entropie de Shannon

$$E(S) = - \sum_{j=1}^{|S|} p(j) \log_2 p(j) \dots\dots\dots (4)$$

Où  $p(j)$  est la probabilité d'avoir un élément de caractéristique  $j$  dans l'ensemble  $S$ .

$$\text{Gain}(S, A) = E(S) - \sum_v \left( \frac{|S_v|}{|S|} * E(S_v) \right) \dots\dots\dots (5)$$

Où  $S$  est un ensemble d'entraînement,  $A$  est l'attribut cible,  $S_v$  le sous-ensemble des éléments dont la valeur de l'attribut  $A$  est  $v$ ,  $|S_v|$  = nombre d'éléments de  $S_v$  et  $|S|$ =nombre d'éléments de  $S$ .

Donc l'idée de base de ID3 est de commencer à construire l'arbre de décision en partant d'un sous ensemble de l'ensemble d'apprentissage, si l'arbre construit classe correctement le reste des données celui-ci sera retenu et l'algorithme se termine, sinon un sous ensemble de données mal classées est ajouté à l'ensemble initial et le processus est réitéré.

- **Les limites de l'algorithme ID3**

L'algorithme ID3 a des limites car il pose quelques problèmes, il ne peut pas traiter les enregistrements incomplets, les attributs sont discrétisés, ce qui n'est pas toujours une solution acceptable, enfin l'arbre produit peut comporter des sous arbres

dans lesquels on ne va presque jamais, Dans ce cadre, il vaut mieux avoir des données exhaustives.

L'algorithme C4.5 permet de répondre à ces limitations de l'algorithme ID3.

## 5.2 L'algorithme C4.5

Cet algorithme été proposé en 1993 aussi par Ross Quinlan, pour pallier les limites de l'algorithme ID3 vu précédemment.

C4.5 repose complètement sur l'algorithme ID3 que nous avons déjà décrit, à partir d'un échantillon d'apprentissage composé d'un certain nombre d'attributs, Ces attributs peuvent être qualitatifs, binaires ou continues, Ainsi, pour les attributs continus, on utilise des heuristiques qui permettent de les discrétiser.

Cet algorithme utilise un critère plus élaboré "le gain ratio" (6) dont le but est de limiter la prolifération de l'arbre en pénalisant les variables qui ont beaucoup de modalités [29].

Le C4.5 fabrique des arbres qui ne sont pas nécessairement binaires (0 à n branches par nœud).

$$\text{GainRatio}(S, A) = \frac{\text{Gain}(S, A)}{\text{SplitInfo}(S, A)} \dots\dots\dots(6)$$

Avec :  $\text{SplitInfo}(S, A) = E(S_v)$

## 5.3 L'algorithme C-RT

L'arbre de décision et de régression (CART) a été conçu par Leo Breiman en 1984, il est considéré comme un arbre binaire puisque chaque nœud ne peut avoir que deux fils, Le principe de l'algorithme de CART s'appuie sur l'indice de Gini (7) pour la segmentation de l'arbre qui vise à construire les feuilles, il faut noter que plus l'indice de Gini est faible, plus le nœud est pur (tous les éléments du nœud appartiennent à la même classe) [30].

$$I_{\text{Gini}} = 1 - \sum_{g=1}^{T_g} (f_g)^2 \dots\dots\dots (7)$$

Telque  $f_g$  la fréquence de la classe g dans le nœud et  $T_g$  le nombre total de classes à prédire [31].

Le principe de cet algorithme consiste à partitionner d'une manière récursive l'ensemble d'entraînement suivant la méthode divisé pour mieux régénérer. En effet, l'algorithme fait une recherche minutieuse, pour chaque nœud, sur les attributs et les valeurs de la segmentation puis il sélectionne le regroupement binaire qui optimise le critère au nœud  $t$ .

## 5.4 L'algorithme Rnd Tree

L'algorithme d'arbre aléatoire fonctionne exactement comme l'arbre de décision avec une exception, pour chaque groupe (split) un sous-ensemble aléatoire d'attributs est disponible, L'opérateur arbre aléatoire fonctionne de manière similaire à C4.5 et C-RT, mais il sélectionne aléatoirement un sous ensemble d'attributs avant qu'il soit appliqué [31].

*"Chacun des petits arbres est moins performant mais l'union fait la force", c'est pour cette raison il faut mettre l'accent sur d'autre sorte d'algorithme se base sur un ensemble d'arbres de décision, c'est la Forêt Aléatoire (RF).*

## 5.5 L'algorithme RF

Les forêts aléatoires ont été introduites par Breiman en 2001, Elles sont en général plus efficaces que les simples arbres de décision.

Une forêt aléatoire est un ensemble d'arbres de décision binaire dans lequel a été introduit de l'aléatoire. Soit  $V_1 \dots V_m$  des variables, une forêt aléatoire est un ensemble de classifieurs  $\{C_i(x, V_i), i = 1 \dots m\}$  où les classifieurs  $C_i$  sont construits sur le même modèle que les arbres binaires. Le nouveau classifieur correspondant à la forêt aléatoire est calculé en prenant la majorité des votes de chacun des classifieurs  $C_i, i = 1 \dots n$  [32].

- ✓ Tous ces algorithmes nous permettent de construire un arbre qui contient des règles pour classer des nouveaux exemples, dans le chapitre suivant on va appliquer ces algorithmes sur notre base de données MMD pour construire le meilleur arbre de décision adaptable à cette base.

## Conclusion

Dans ce chapitre on a présenté en premier temps les SADM, ensuite on a parlé de la classification qui est en relation direct avec ce type de système, en deuxième temps, on a choisi la base de données "Mammographic Mass Data" pour extraire la connaissance qu'on a besoin, cette base de données est très adaptable à la problématique posée dans le premier chapitre, selon la constitution de cette base de données la classification basé sur l'apprentissage supervisé est la classification choisie pour réaliser notre travaille, ensuite on a mis l'accent sur l'étude de l'état de l'art concernant l'utilisation de cette base de données afin de fixer les méthodes de classification à utiliser dans le chapitre suivant, notre choix est tombé sur les arbres de décision.

Dans le chapitre suivant on va appliquer les arbres de décision sur la base de données MMD afin d'obtenir un classifieur qui nous permet d'avoir des bons résultats.

# Chapitre 3

Expérimentation et réalisation de l'application



## Introduction

Dans les chapitres précédents on a spécifié la problématique principale à résoudre dans ce projet, c'est l'amélioration de la classification BIRADS, pour ce fait on a choisi la base de données MMD pour réaliser notre travail, on a trouvé que cette BDD a besoin d'être prétraité avant de l'utiliser.

Dans ce chapitre, en premier temps on va essayer de traiter les données manquantes pour avoir une BDD équilibré et complète en utilisant différentes approches, en deuxième temps on va appliquer la classification sur notre BDD en utilisant différent algorithmes afin d'extraire un modèle (arbre) qui nous permet de classer des nouveaux cas, finalement, dans la dernière partie, on va réaliser notre application par l'implémentation de l'arbre optimal, pour que notre modèle être utilisable par un praticien dans la détection de type d'une masse mammaire.

La procédure suivie dans ce chapitre est représentée dans le schéma suivant :



Figure 3.1 - Représentation de la procédure suivie dans le chapitre 3.

### 1 Prétraitement de la base de données

La BDD qu'on a choisi dans le chapitre précédent et qui est adaptable à notre problématique est MMD qui contient 961 exemples avec 5 attributs (descripteurs) et la classe de sortie, Dans le cadre d'étude de cette BDD on a trouvé qu'elle contient des données manquantes, qui sont les suivantes:

- BIRADS : 2 données manquantes
- Age : 5
- Forme : 31
- Marge (contour) : 48
- Densité: 76

## 1.1 Méthodes pour traiter les données manquantes

Face à la présence de DM, il existe différentes méthodes pour résoudre ce problème, parmi ces méthodes on distingue:

1. La suppression des enregistrements c'est-à-dire exclure du fichier de données tous les individus ayant au moins une DM, mais avec cette méthode on va réduire la fiabilité de la BDD.
2. L'imputation simple qui consiste à remplacer chaque donnée manquante par une valeur plausible, par exemple, remplacer toutes les DM par la moyenne calculée sur les données réellement observées, ou par la variable 0 ou 1, mais cette méthode influence sur l'intégrité de la BDD.
3. D'autres méthodes d'imputation simple sont également disponibles, comme l'imputation par le plus proche voisin (méthode de classification) qui remplace les données manquantes par des valeurs provenant d'individus similaires pour lesquels toute l'information a été observée, et l'imputation par régression qui consiste à remplacer les DM par des valeurs prédites selon un modèle de régression [33].

La prédiction des DM par la troisième méthode est donc une solution valide, car la prédiction de la donnée ici est par l'utilisation d'un modèle qui va étudier une BDD afin de prédire les données manquantes en fonction des autres données.

A la suite, on va traiter les DM par la prédiction en utilisant la classification par arbre de décision.

## 1.2 Etapes de prétraitement

1. Filtrer la BDD, en gardant juste les exemples qui n'ont pas de DM, après le filtrage on obtient:

- Une base de données MMD1 qui contient 830 exemples.
- Cinq petites bases de données, chaque une contient un ensemble d'exemples avec un attribut à prédire, par exemple une base qui contient l'attribut "Densité" comme donnée manquante, donc notre but est de prédire les valeurs de cet attribut en utilisant les autres descripteurs et un modèle de classification.

2. On utilise la base de données MMD1, définir comme classe de sortie l'attribut Densité, puis on applique la classification sur cette BDD afin d'obtenir un classifieur optimal qui nous permet de prédire les valeurs de donnée manquante "Densité".
3. Après la prédiction de la "Densité" par la classification, on ajoute les exemples à la base de données MMD1, on obtient une base de données MMD2.
4. Chaque fois, on utilise la nouvelle base de données et on répète les étapes 2 et 3 pour chaque attribut, jusqu'à prédire tous les DM, et finalement, on obtient une BDD qui est prête à utiliser.

## 1.3 Travail effectué

### 1.3.1 L'échantillonnage de la base de données

Dans la littérature, il existe différentes méthodes pour l'échantillonnage, parmi ces méthodes on distingue la validation croisée et la méthode de plus grande portion pour l'apprentissage et le reste pour le test, on va utiliser cette dernière dans la partie de prétraitement des DM.

On a essayé de travailler par les valeurs (70%, 30%) pour (apprentissage/test) et aussi par les valeurs (80%, 20%), cette dernière nous donne des bons résultats que la première, donc, à la suite de cette partie on va travailler par l'échantillonnage:

{ 80% pour l'apprentissage  
20% pour les testes

### 1.3.2 Classifieurs utilisés

Notre but est de construire un arbre de décision pour prédire les données manquantes, pour la construction de l'arbre on va utiliser différents algorithmes qui sont disponibles sous Tanagra, telque : C4.5, ID3 et C-RT.

### 1.3.3 Résultats

#### A. Prédiction de la Densité

On a utilisé la base de données MMD1, en prenant comme classe de sortie l'attribut Densité, puis on a appliqué la classification sur cette base de données, enfin et après plusieurs tests on a trouvé les résultats représenté dans le tableau 3.1.

Algorithme	Taux de classification
C4.5	90.81%
ID3	90.66%
C-RT	90.66%

Tableau 3.1 - TC obtenu par chaque algorithme pour l'attribut Densité.

- *Discutions*

D'après le Tableau 3.1, il est remarquable que l'algorithme C4.5 donne le meilleur TC qui est 90.81%, donc on prend l'arbre qui est construit par cet algorithme (Figure 3.2) pour prédire la valeur de l'attribut Densité.

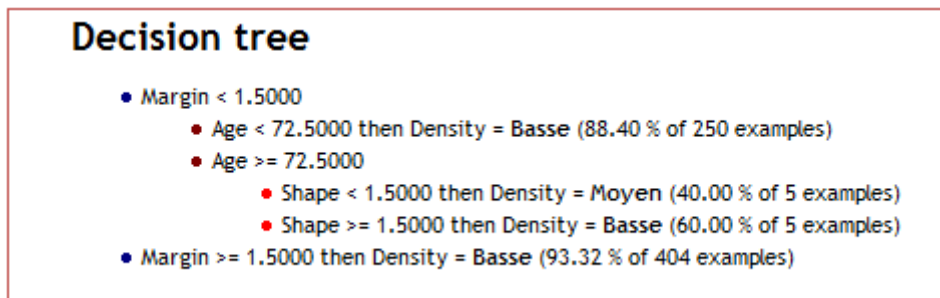


Figure 3.2 - Arbre construit par C4.5 pour prédire la valeur de l'attribut Densité.

- *Implémentation de l'arbre sous C++*

Pour utiliser l'arbre dans la prédiction des données, on se base sur l'extraction des règles de classification à partir de cet arbre, on a choisi le langage C++ pour implémenter ces règles afin de classer les exemples qui ont "Densité" comme donnée manquantes.

## B. Prédiction de BIRADS, Age, Forme et Contour

On applique la même procédure précédente, pour prédire les données manquantes pour BIRADS, Age, Forme et Contour.

- **BIRADS**

Algorithme	Taux de classification
C4.5	77.94%
ID3	75.59%
C-RT	75.59%

Tableau 3.2 - TC obtenu par chaque algorithme pour l'attribut BIRADS.

- ✓ L'algorithme C4.5 donne le meilleur taux de classification qui est 77.94%.
- ✓ L'arbre construit pour prédire des nouveaux cas est représenté dans la Figure 3.3.

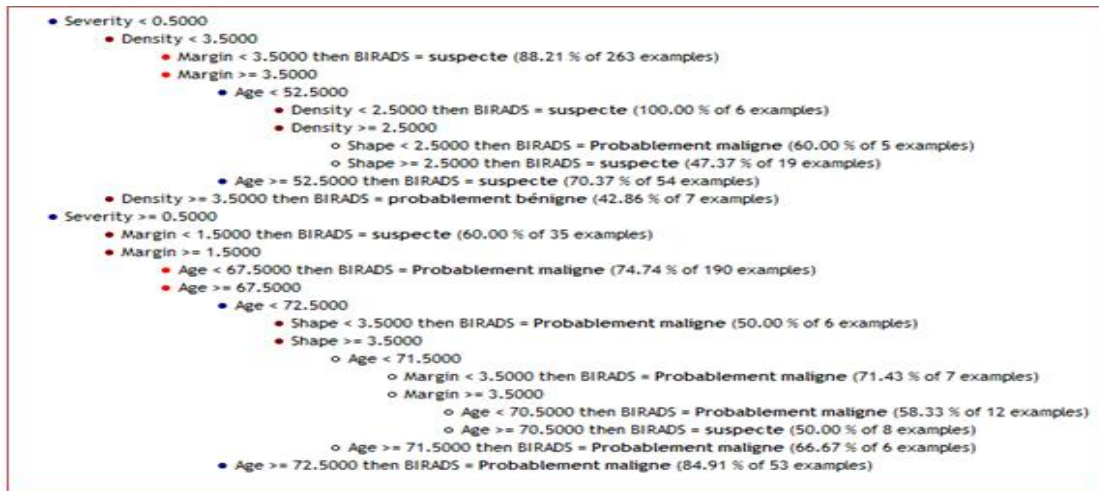


Figure 3.3 - Arbre construit par C4.5 pour prédire la valeur de l'attribut BIRADS.

- **Age**

Algorithme	Taux de classification
C4.5	62.26%
ID3	55.47%
C-RT	53.99%

Tableau 3.3 - TC obtenu par chaque algorithme pour l'attribut Age.

- ✓ L'algorithme C4.5 donne le meilleur taux de classification qui est 62.26%.
- ✓ Lorsque l'arbre construit est très grand (Annexe A), on représente juste une description (Figure 3.4), cet arbre contient 67 nœuds et 34 règles.

Tree description	
Number of nodes	67
Number of leaves	34

Figure 3.4 - Description de l'arbre construit par C4.5 pour prédire la valeur de l'attribut Age.

- **Forme**

Algorithme	Taux de classification
C4.5	67.91%
ID3	63.50%
C-RT	63.50%

Tableau 3.4 - TC obtenu par chaque algorithme pour l'attribut Forme.

- ✓ L'algorithme C4.5 donne le meilleur taux de classification qui est 67.91%.
- ✓ Lorsque l'arbre construit est très grand (Annexe B), on représente juste une description (Figure 3.5).

Tree description	
Number of nodes	41
Number of leaves	21

Figure 3.5 - Description de l'arbre construit par C4.5 pour prédire la valeur de l'attribut Forme.

- **Contour**

Algorithme	Taux de classification
C4.5	67.64%
ID3	62.48%
C-RT	62.48%

Tableau 3.5 - TC obtenu par chaque algorithme pour l'attribut Contour.

- ✓ L'algorithme C4.5 donne le meilleur taux de classification qui est 67.64%.

**Remarque**

On remarque que les meilleurs TC sont obtenus généralement par l'algorithme C4.5, et ça ce que vérifier la récente liste de classement des classifieurs (Top Ten) ou le C4.5 est le premier [34].

- ✓ Lorsque l'arbre construit est très grand (Annexe C), on représente juste une description (Figure 3.6).

Tree description	
Number of nodes	39
Number of leaves	20

Figure 3.6 - Description de l'arbre construit par C4.5 pour prédire la valeur de l'attribut Contour.

Pour chaque attribut on a créé un programme C++, qui demande à l'utilisateur de saisir les descripteurs disponible concernant un exemple, pour les utiliser dans la classification en utilisant les règles qui sont programmées dans ce programme, enfin on a rassemblé tous ces programmes dans une application C++ (PDM) (Figure 3.7), pour faciliter leur utilisation dans la prédiction des données manquantes.



Figure 3.7 - Interface de l'application PDM.

- *Exemple d'exécution de l'application PDM*

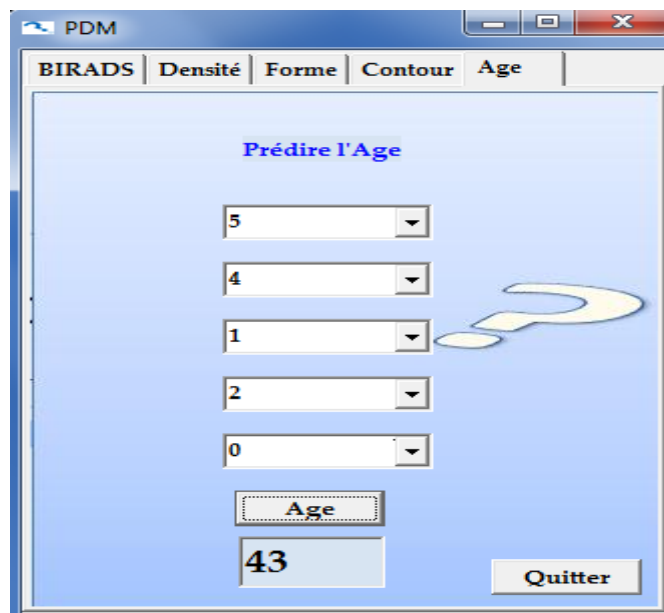


Figure 3.8 - Exécution de l'application PDM.



Après l'utilisation de notre application PDM, on a calculé la valeur de chaque donnée manquante, et finalement on a une base de données complète qui contient 961 exemples avec 5 descripteurs et une classe de sortie, cette base de données est prête à utiliser dans la phase suivante.

## **2 Extraction de l'arbre optimal**

### **2.1 Classifieurs utilisés**

Dans le chapitre précédent on a étudié l'état de l'art concernant l'utilisation de la base de données MMD, et on a trouvé qu'elle est utilisée par plusieurs chercheurs, enfin notre choix est tombé sur l'utilisation des arbres de décision selon différents arguments.

Pour la construction de l'arbre de décision, on va appliquer différents algorithmes sur notre base de données qui est prête à utiliser, tel que C4.5, ID3 et C-RT.

### **2.2 Echantillonnage utilisé**

On va travailler par deux méthodes, la méthode classique c'est-à-dire la plus grande partie pour l'apprentissage et le reste pour le test (on va essayer avec 80% et 20% puis 70% et 30%), et la méthode de la validation croisée (on va utiliser 10 Partitions), afin de comparer les résultats obtenus et avoir la meilleure méthode adaptable à la base de données MMD.

### **2.3 Travail effectué**

#### **2.3.1 Classification par l'algorithme C4.5**

Par l'application de l'algorithme C4.5 sur la base de données MMD, et après plusieurs tests (changement des paramètres de cet algorithme) on obtient les résultats représentés dans le Tableau 3.6.

Partition	Meilleur TC
70% / 30%	88.84%
80% / 20%	89.07%
Validation croisé (10 partitions)	83.44%

Tableau 3.6 - TC obtenu par l'algorithme C4.5 en utilisant différentes partition.

### 2.3.2 Classification par l'algorithme ID3

Par l'application de l'algorithme ID3 sur la base de données MMD, et après plusieurs tests (changement des paramètres de cet algorithme) on a obtenu les résultats représentés dans le Tableau 3.7.

Partition	Meilleur TC
70% / 30%	87.35%
80% / 20%	87.11%
Validation croisé (10 partitions)	81.77%

Tableau 3.7 - TC obtenu par l'algorithme ID3 en utilisant différentes partition.

### 2.3.3 Classification par l'algorithme C-RT

Par l'application de l'algorithme C-RT sur la base de données MMD, et après plusieurs tests (changement des paramètres de cet algorithme) on a obtenu les résultats représentés dans le Tableau 3.8.

Partition	Meilleur TC
70% / 30%	85.71%
80% / 20%	86.59%
Validation croisé (10 partitions)	81.98%

Tableau 3.8 - TC obtenu par l'algorithme C-RT en utilisant différentes partition.

- **Discussion**

D'après le Tableau 3.6, le meilleur TC par l'utilisation de l'algorithme C4.5 égale à 89.07% en utilisant l'échantillonnage 80% pour l'apprentissage et 20% pour le test.

D'après le Tableau 3.7, le meilleur TC par l'utilisation de l'algorithme ID3 égale à 87.35% en utilisant l'échantillonnage 70% pour l'apprentissage et 30% pour le test.

Selon le Tableau 3.8, le meilleur taux de classification par l'utilisation de l'algorithme C-RT égale à 86.59% en utilisant l'échantillonnage 80% pour l'apprentissage et 20% pour le test.

### 2.3.4 Comparaison des Résultats

On résume les meilleurs résultats obtenus dans un histogramme (Figure 3.9).

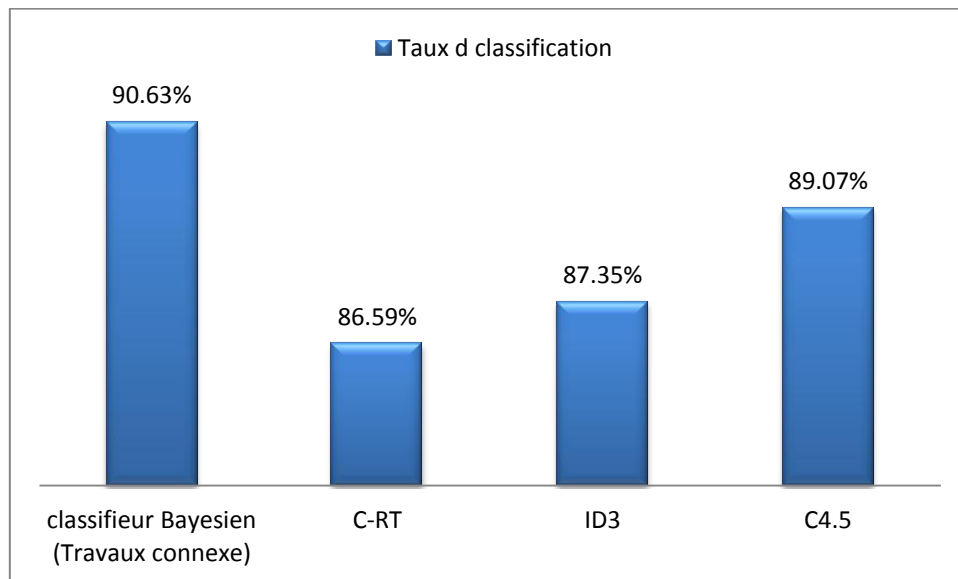


Figure 3.9 - Histogramme représente le TC obtenu par chaque algorithme.

- **Discussion**

Selon l'histogramme au-dessus il est remarquable que l'algorithme C4.5 donne le meilleur TC (89.07%) dans notre travail.

Le meilleur TC trouvé durant l'étude de l'état de l'art égale à 90.63%, et ce résultat est meilleur que celui trouvé dans notre travail (89.07%), c'est pour cette raison on

passé à l'utilisation des autres types d'algorithmes qui se base sur l'aléatoire pour trouver des bons résultats.

A la suite on va appliquer les algorithmes (Rnd Tree) et RF sur la base de données MMD, pour améliorer les résultats trouvés précédemment.

### 2.3.5 Classification par l'algorithme Rnd Tree

Par l'application de l'algorithme Rnd Tree sur la base de données MMD, et après plusieurs tests (changement des paramètres de cet algorithme) on a obtenu les résultats représentés dans le Tableau 3.9.

Partition	Meilleur TC
70% / 30%	94.35%
80% / 20%	93.75%
Validation croisé (10 partitions)	77.71%

Tableau 3.9 - TC obtenu par l'algorithme Rnd Tree en utilisant différentes partition.

### 2.3.6 Classification par l'algorithme de RF

Par l'application de l'algorithme RF sur la base de données MMD, et après plusieurs tests (changement de nombre d'arbres) on obtient les résultats représentés dans la figure 3.10.

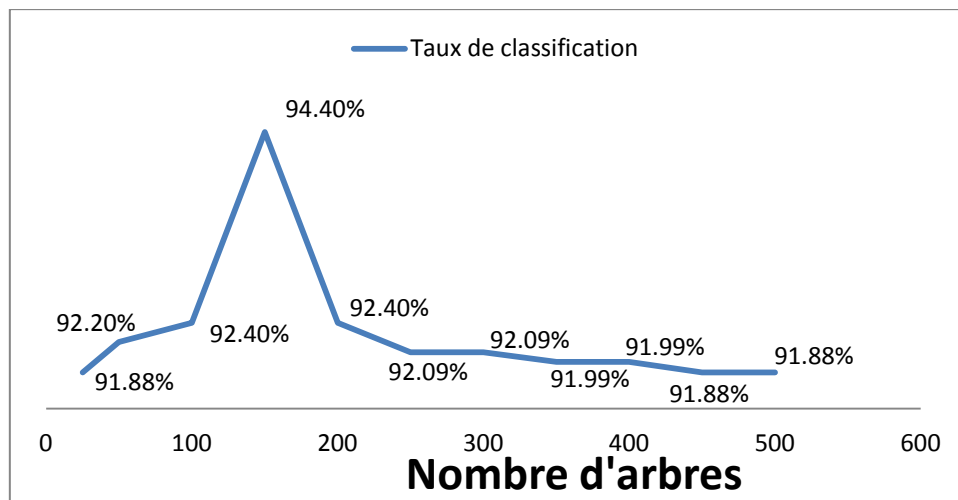


Figure 3.10 - Courbe représente le TC obtenu en fonction de nombre d'arbre par RF.

- **Discussion**

D'après le Tableau 3.9, le meilleur TC par l'utilisation de l'algorithme Rnd Tree égale à 94.35% en utilisant l'échantillonnage 70% pour l'apprentissage et 30% pour le test.

D'après la courbe (Figure 3.10) on remarque qu'il y a une augmentation de TC jusqu'à atteint un seuil maximal (TC = 94.40%) par 150 arbres, après on a une diminution de TC en augmentant le nombre d'arbre, à partir de 450 arbres on remarque qu'il y a une stabilisation de TC, alors, Le meilleur TC par l'utilisation de l'algorithme RF égale à 94.4% en utilisant 150 arbres (de type Rnd Tree).

- **Comparaison**

Selon les résultats obtenus dans les deux dernières expériences on remarque que le meilleur TC par Rnd Tree égale à 94.35%, et par RF égale à 94.4%, donc les résultats sont très proches.

Ces deux résultats sont meilleurs à celle trouvé dans les travaux connexe (90. 63%), alors on a une amélioration de TC durant notre travaille.

Lorsque les résultats obtenus sont très proches, on passe à un autre critère de comparaison, c'est le temps de calcule (Figure 3.11).

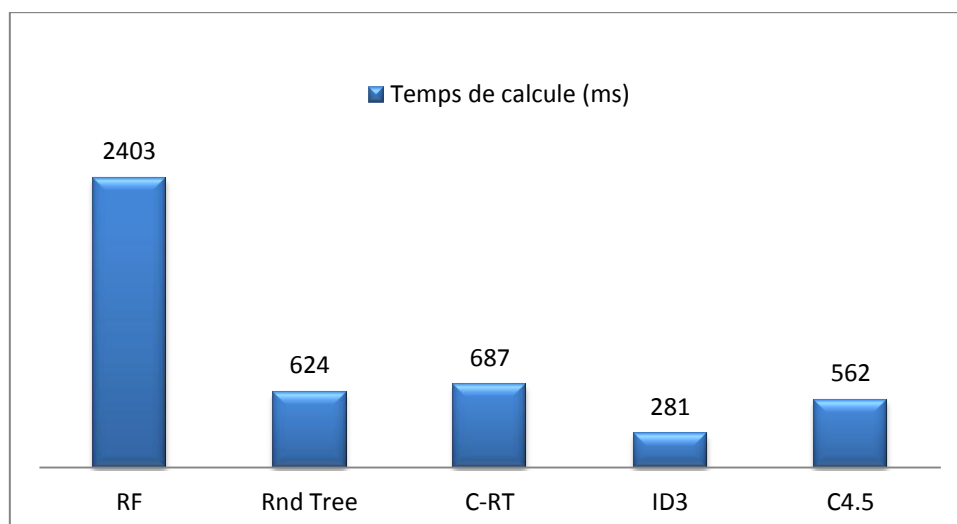


Figure 3.11 - Histogramme représente le temps de calcule par chaque algorithme pour donner le meilleur TC.

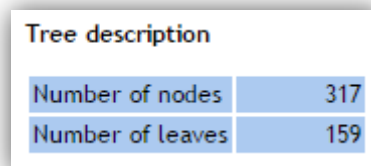
- *Discussion*

Le meilleur TC égale à 94.4% par RF, mais on remarque selon l'histogramme que cet algorithme prend la plus grande durée de calcul pour donner ce résultat, car l'algorithme utilise plusieurs arbres donc un temps de calcul important.

L'algorithme qui donne un TC meilleur à celle trouvé dans les travaux connexe et dans un temps acceptable (624ms) c'est Rnd Tree.

## 2.4 Résultat Final

Notre arbre optimal (Annexe D) pour classer des nouveaux cas est l'arbre qui est construit par l'algorithme Rnd Tree (Figure 3.12).



Tree description	
Number of nodes	317
Number of leaves	159

Figure 3.12 - Description de l'arbre optimal.

Selon la description de l'arbre optimal on remarque qu'il contient 159 nœud feuille, alors on peut extraire 159 règles à partir de cet arbre.

## 3 Réalisation de l'application en implémentant l'arbre optimal

Dans la partie précédente, on a essayé d'appliquer différents algorithmes sur la base de données MMD afin d'extraire l'arbre optimal qui peut classer des nouveaux cas de la meilleure façon.

On a obtenu comme résultat final un arbre avec un TC égale à 94.35%, cet arbre peut être transformé à un ensemble de règles (159 règles).

Notre but est d'avoir une petite application qui est facile à utiliser par un praticien dans la classification de type d'une masse mammaire après la classification BIRADS, donc on a implémenté l'ensemble de règles obtenu par notre arbre optimal en utilisant

le langage C++, puis on a intégré cette implémentation avec une simple interface pour être utilisable (Figure 3.13).



Figure 3.13 - Interface finale de l'application.

### 3.1 Description

- Le nom de notre application est DTMM, on a choisi comme icône pour cette application le ruban rose pour indiquer que cette application est en relation directe avec le cancer de sein (pour détecter le type d'une masse mammaire) (Figure 3.14).



Figure 3.14 - Icône de l'application DTMM.

L'interface de l'application DTMM contient deux fenêtres, une fenêtre pour afficher le résultat et une fenêtre pour l'aide.

- La fenêtre "Résultat" contient trois boutons, le premier pour charger l'image à étudier, le deuxième pour afficher le résultat finale de la classification, ce bouton se base sur notre ensemble de règles, et finalement le troisième bouton pour quitter l'application, On a aussi deux champs texte, l'un pour saisir l'âge de la patiente et l'autre pour afficher le résultat final.

Lorsque les descripteurs qu'il faut saisir sont des descripteurs catégoriques, on a choisi d'utiliser la notion de listes, donc on a dans notre interface quatre champs de type "ComboBox" pour choisir la valeur de BIRADS, Forme, Contour et Densité à partir de l'image à étudier.

- La fenêtre "Aide" est une fenêtre supplémentaire pour aider le praticien dans le choix de différentes catégories de Forme, Contour et Densité (Figure 3.15) s'il y a une complexité, c'est un ensemble d'image et chaque image représente un exemple de la catégorie de l'attribut étudié.

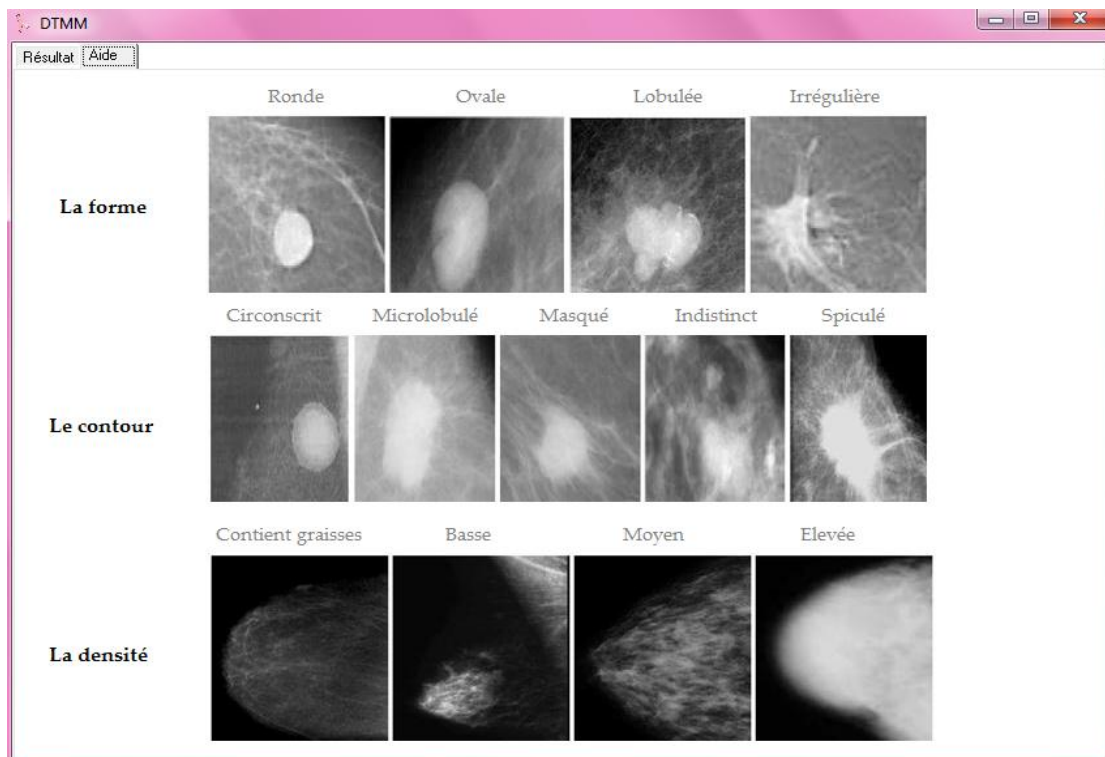


Figure 3.15 - La fenêtre Aide de l'application DTMM.



## 3.2 Exécution

Lorsque le praticien exécute notre application la première des choses c'est de charger l'image à étudier, puis il va saisir l'âge de la patiente et le résultat du rapport de la classification BIRADS, ensuite il va passer à l'étude de l'image affichée déjà, il va choisir la forme, le contour et la densité de la masse (s'il y a des complexités durant le choix, il va utiliser la fenêtre Aide), finalement, lorsque le praticien clic sur le bouton Résultat, il va obtenir le résultat finale, c'est le type de la masse mammaire soit bénigne, soit maligne (Figure 3.16).

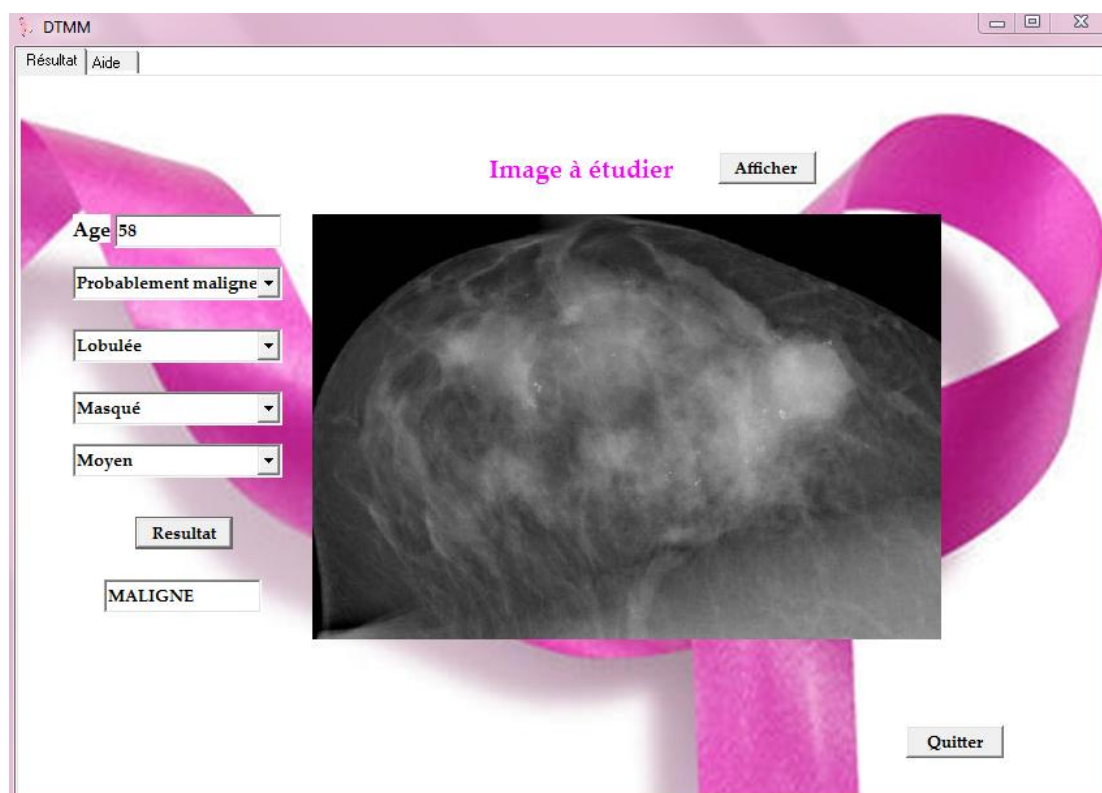


Figure 3.16 - Exécution de l'application.

## Conclusion

Dans ce dernier chapitre, on a traité les données manquantes de la base de données MMD par l'utilisation des arbres de décision, les meilleurs résultats sont obtenus par l'algorithme C4.5 , à la suite on a appliqué différents algorithmes sur notre base afin d'extraire un arbre optimal, l'algorithme qui nous donne l'arbre optimal est Rnd Tree, par notre propre démarche on a trouvé des bons résultats par rapport à celle trouvée dans les travaux connexes, enfin on a transformé l'arbre trouvé à un ensemble de règles et intégré ce dernier avec une simple interface pour être facile à utiliser dans la classification des nouveaux cas.

---

# Conclusion générale

Nous nous sommes intéressés dans ce travail à réaliser une application pour la détection de type d'une masse mammaire dans le cadre d'aide au diagnostic et décision médical, en améliorant la classification BIRADS.

A cet effet, nous avons procédé en premier temps à une analyse de travaux effectués sur la même problématique qui a permis de mettre en évidence les outils et les méthodes à utiliser, enfin notre choix est d'utiliser la classification basé sur l'apprentissage supervisé et plus particulièrement les arbres de décision.

Les arbres de décision présentent un avantage majeur dans la classification par leur simplicité, et leur facilité d'interprétation. L'induction des règles de décision à partir de l'arbre induit représente l'un de ses avantages principaux.

Pour obtenir l'arbre optimal, nous avons appliqué différents algorithmes (C4.5, ID3, C-RT, Rnd Tree et RF) sur la base de données MMD de l'UCI qui est très adaptable à notre problématique.

Nous avons évalué et testé les performances de ces algorithmes en termes de taux de classification (TC), les meilleurs résultats sont obtenus par l'algorithme Rnd Tree. Le taux de classification obtenu avec notre méthode reste le meilleur résultat obtenu jusqu'à maintenant pour la classification de la base de données MMD, alors par notre propre démarche on a une amélioration des travaux précédents dans ce sujet.

Notons que la méthode que nous présentons dans ce mémoire répond à un besoin vital dans le domaine médical et offre aux médecins une base de connaissance explicite (sous forme de règles) acquise d'une base de données médicale. L'expert aura la possibilité d'accepter les règles, de les modifier, de les supprimer ou d'ajouter d'autres.

Pour faciliter l'utilisation de ces règles dans la classification de type d'une masse mammaire, nous avons les implémenter et les intégrer avec une simple application.

Dans les perspectives de ce travail, nous souhaitons apporter d'autres améliorations aux arbres de décisions en utilisant la notion d'hybridation. Nous envisageons également intégrer cette application dans un système d'aide au diagnostic médical pour l'utilisation dans un hôpital ou dans un centre d'imagerie.

# Annexes

## Annexe A : Arbre construit par C4.5 pour prédire les valeurs de l'attribut Age.

Margin < 2.5000

BIRADS < 4.5000

Margin < 1.5000

Density < 2.5000

Shape < 1.5000 then Age = quarante trois

Shape >= 1.5000 then Age = quarante neuf

Density >= 2.5000

Severity < 0.5000

BIRADS < 3.5000 then Age = cinquante sept

BIRADS >= 3.5000

Shape < 1.5000 then Age = soixante quatre

Shape >= 1.5000 then Age = quarante et un

Severity >= 0.5000 then Age = soixante dix

Margin >= 1.5000 then Age = quarante et un

BIRADS >= 4.5000

Shape < 2.5000 then Age = quarante trois

Shape >= 2.5000 then Age = soixante douze

Margin >= 2.5000

Density < 2.5000

BIRADS < 4.5000 then Age = soixante

BIRADS >= 4.5000 then Age = cinquante neuf

Density >= 2.5000

Density < 3.5000

Shape < 2.5000

Shape < 1.5000

Severity < 0.5000 then Age = soixante

Severity >= 0.5000

Margin < 4.5000 then Age = quarante trois

Margin >= 4.5000 then Age = cinquante sept

Shape >= 1.5000

Margin < 3.5000 then Age = trente six

Margin >= 3.5000

Severity < 0.5000 then Age = cinquante sept

Severity >= 0.5000 then Age = soixante

Shape >= 2.5000

BIRADS < 4.5000

BIRADS < 3.5000 then Age = soixante douze

BIRADS >= 3.5000

Shape < 3.5000

Severity < 0.5000

Margin < 3.5000 then Age = soixante deux

Margin >= 3.5000 then Age = cinquante et un

Severity >= 0.5000 then Age = cinquante cinq

Shape >= 3.5000

Severity < 0.5000

Margin < 3.5000 then Age = soixante six

Margin >= 3.5000

Margin < 4.5000 then Age = cinquante six

Margin >= 4.5000 then Age = soixante sept

Severity >= 0.5000

Margin < 4.5000 then Age = cinquante sept

Margin >= 4.5000 then Age = cinquante

BIRADS >= 4.5000

BIRADS < 5.5000

Shape < 3.5000 then Age = soixante cinq

Shape >= 3.5000

Severity < 0.5000

Margin < 4.5000 then Age = cinquante sept  
 Margin >= 4.5000 then Age = quarante deux  
 Severity >= 0.5000  
 Margin < 4.5000  
 Margin < 3.5000 then Age = soixante huit  
 Margin >= 3.5000 then Age = soixante six  
 Margin >= 4.5000 then Age = soixante sept  
 BIRADS >= 5.5000 then Age = soixante seize  
 Density >= 3.5000 then Age = cinquante\_ sept

### **Annexe B : Arbre construit par C4.5 pour prédire les valeurs de l'attribut Forme.**

Margin < 1.5000  
 Density < 1.5000 then Shape = Ronde  
 Density >= 1.5000  
 Age < 20.5000 then Shape = Ronde  
 Age >= 20.5000  
 Age < 68.5000  
 Density < 2.5000 then Shape = Ovale  
 Density >= 2.5000  
 Age < 66.5000  
 Age < 28.5000 then Shape = Ovale  
 Age >= 28.5000 then Shape = Ronde  
 Age >= 66.5000 then Shape = Ronde  
 Age >= 68.5000 then Shape = Ovale  
 Margin >= 1.5000  
 Density < 1.5000 then Shape = Irrégulière  
 Density >= 1.5000  
 Severity < 0.5000  
 Margin < 4.5000  
 BIRADS < 3.5000 then Shape = Irrégulière  
 BIRADS >= 3.5000  
 Margin < 2.5000 then Shape = Irrégulière

Margin >= 2.5000  
 Age < 45.0000  
 Margin < 3.5000 then Shape = Ovale  
 Margin >= 3.5000 then Shape = Lobulée  
 Age >= 45.0000  
 Age < 59.5000  
 Age < 49.5000 then Shape = Irrégulière  
 Age >= 49.5000  
 Age < 53.5000 then Shape = Lobulée  
 Age >= 53.5000 then Shape = Irrégulière  
 Age >= 59.5000  
 Age < 63.5000 then Shape = Lobulée  
 Age >= 63.5000  
 Age < 70.5000 then Shape = Irrégulière  
 Age >= 70.5000  
 Margin < 3.5000 then Shape = Ovale  
 Margin >= 3.5000 then Shape = Lobulée  
 Margin >= 4.5000 then Shape = Irrégulière  
 Severity >= 0.5000 then Shape = Irrégulière

### **Annexe C : Arbre construit par C4.5 pour prédire les valeurs de l'attribut Contour.**

Shape < 2.5000  
 BIRADS < 4.5000 then Margin = Circonscrit  
 BIRADS >= 4.5000 then Margin = Indistinct  
 Shape >= 2.5000  
 Age < 29.5000 then Margin = Circonscrit  
 Age >= 29.5000  
 Shape < 3.5000  
 Age < 75.0000  
 BIRADS < 4.5000  
 Age < 61.5000 then Margin = Indistinct

Age  $\geq 61.5000$  then Margin = Masqué  
 BIRADS  $\geq 4.5000$  then Margin = Indistinct

Age  $\geq 75.0000$  then Margin = Indistinct

Shape  $\geq 3.5000$

Age  $< 46.5000$

Age  $< 38.0000$  then Margin = Indistinct

Age  $\geq 38.0000$

BIRADS  $< 4.5000$  then Margin = Indistinct

BIRADS  $\geq 4.5000$  then Margin = Spiculé

Age  $\geq 46.5000$

Density  $< 2.5000$  then Margin = Indistinct

Density  $\geq 2.5000$

Age  $< 49.5000$

Age  $< 48.5000$  then Margin = Spiculé

Age  $\geq 48.5000$  then Margin = Indistinct

Age  $\geq 49.5000$

BIRADS  $< 4.5000$

Age  $< 51.5000$  then Margin = Spiculé

Age  $\geq 51.5000$

BIRADS  $< 3.5000$  then Margin = Masqué

BIRADS  $\geq 3.5000$  then Margin = Indistinct

BIRADS  $\geq 4.5000$

BIRADS  $< 5.5000$

Age  $< 86.5000$

Age  $< 56.5000$  then Margin = Spiculé

Age  $\geq 56.5000$  then Margin = Indistinct

Age  $\geq 86.5000$  then Margin = Spiculé

BIRADS  $\geq 5.5000$  then Margin = Indistinct

### **Annexe D : L'arbre optimal construit par Rnd Tree.**

BIRADS  $< 4.5000$

Shape  $< 3.5000$

Age  $< 40.5000$  then Severity = bénigne

Age  $\geq 40.5000$

Density  $< 1.5000$

Age  $< 44.0000$  then Severity = maligne

Age  $\geq 44.0000$

Age  $< 72.0000$  then Severity = bénigne

Age  $\geq 72.0000$

Age  $< 76.0000$  then Severity = maligne

Age  $\geq 76.0000$  then Severity = bénigne

Density  $\geq 1.5000$

Density  $< 2.5000$  then Severity = bénigne

Density  $\geq 2.5000$

Age  $< 66.5000$

Shape  $< 2.5000$

Age  $< 44.5000$

Margin  $< 1.5000$

Density  $< 3.5000$

Shape  $< 1.5000$

BIRADS  $< 3.5000$  then Severity = bénigne

BIRADS  $\geq 3.5000$

Age  $< 42.5000$

Age  $< 41.5000$  then Severity = bénigne

Age  $\geq 41.5000$  then Severity = bénigne

Age  $\geq 42.5000$

Age  $< 43.5000$  then Severity = bénigne

Age  $\geq 43.5000$  then Severity = maligne

Shape  $\geq 1.5000$

Age  $< 42.5000$

BIRADS  $< 3.5000$  then Severity = maligne

BIRADS  $\geq 3.5000$

Age  $< 41.5000$  then Severity = bénigne

Age  $\geq 41.5000$  then Severity = maligne

---

Age >= 42.5000 then Severity = bénigne	Age < 63.0000
Density >= 3.5000 then Severity = bénigne	Age < 46.5000
Margin >= 1.5000 then Severity = bénigne	Age < 44.0000 then Severity = bénigne
Age >= 44.5000	Age >= 44.0000 then Severity = maligne
Margin < 1.5000	Age >= 46.5000
Shape < 1.5000	Age < 55.0000 then Severity = bénigne
Age < 52.5000 then Severity = bénigne	Age >= 55.0000
Age >= 52.5000	Age < 56.5000 then Severity = maligne
Age < 54.5000	Age >= 56.5000
Age < 53.5000 then Severity = bénigne	Age < 61.5000
Age >= 53.5000 then Severity = bénigne	Margin < 3.5000 then Severity = maligne
Age >= 54.5000	Margin >= 3.5000
Age < 65.5000	Age < 60.5000 then Severity = bénigne
Age < 58.5000	Age >= 60.5000 then Severity = maligne
Age < 57.5000 then Severity = bénigne	Age >= 61.5000 then Severity = bénigne
Age >= 57.5000 then Severity = bénigne	Age >= 63.0000
Age >= 58.5000 then Severity = bénigne	Age < 64.5000 then Severity = maligne
Age >= 65.5000	Age >= 64.5000
BIRADS < 3.0000 then Severity = bénigne	Age < 65.5000 then Severity = maligne
BIRADS >= 3.0000 then Severity = bénigne	Age >= 65.5000 then Severity = bénigne
Shape >= 1.5000	Age >= 66.5000
Age < 45.5000 then Severity = bénigne	Margin < 3.5000
Age >= 45.5000 then Severity = bénigne	BIRADS < 3.5000 then Severity = maligne
Margin >= 1.5000	BIRADS >= 3.5000
Age < 52.5000 then Severity = bénigne	Margin < 1.5000
Age >= 52.5000	Age < 73.5000
Margin < 2.5000 then Severity = maligne	Age < 68.5000
Margin >= 2.5000	Age < 67.5000 then Severity = bénigne
Age < 58.0000 then Severity = bénigne	Age >= 67.5000 then Severity = maligne
Age >= 58.0000	Age >= 68.5000 then Severity = bénigne
Margin < 3.5000 then Severity = bénigne	Age >= 73.5000 then Severity = maligne
Margin >= 3.5000 then Severity = maligne	Margin >= 1.5000 then Severity = bénigne
Shape >= 2.5000	



---

Margin $\geq 3.5000$ then Severity = maligne	Age $< 49.5000$
Shape $\geq 3.5000$	BIRADS $< 3.5000$ then Severity = bénigne
Age $< 47.5000$	BIRADS $\geq 3.5000$ then Severity = maligne
Density $< 2.5000$	Age $\geq 49.5000$
Age $< 43.5000$ then Severity = maligne	BIRADS $< 3.5000$ then Severity = maligne
Age $\geq 43.5000$ then Severity = bénigne	BIRADS $\geq 3.5000$
Density $\geq 2.5000$ then Severity = bénigne	Age $< 51.0000$
Age $\geq 47.5000$	Margin $< 4.5000$ then Severity = maligne
Density $< 3.5000$	Margin $\geq 4.5000$ then Severity = maligne
Age $< 69.5000$	Age $\geq 51.0000$
Margin $< 3.5000$	Age $< 53.5000$
Age $< 64.5000$	Age $< 52.5000$ then Severity = maligne
Age $< 58.0000$	Age $\geq 52.5000$ then Severity = bénigne
Age $< 56.0000$	Age $\geq 53.5000$
Age $< 50.5000$	Age $< 54.5000$ then Severity = maligne
BIRADS $< 3.0000$ then Severity = bénigne	Age $\geq 54.5000$
BIRADS $\geq 3.0000$ then Severity = maligne	Age $< 56.5000$ then Severity = bénigne
Age $\geq 50.5000$ then Severity = bénigne	Age $\geq 56.5000$
Age $\geq 56.0000$ then Severity = maligne	Age $< 63.5000$
Age $\geq 58.0000$ then Severity = bénigne	Age $< 60.5000$
Age $\geq 64.5000$ then Severity = maligne	Margin $< 4.5000$
Margin $\geq 3.5000$	Age $< 58.5000$
Density $< 2.5000$ then Severity = maligne	Age $< 57.5000$ then Severity = maligne
Density $\geq 2.5000$	Age $\geq 57.5000$ then Severity = maligne
BIRADS $< 2.5000$	Age $\geq 58.5000$ then Severity = maligne
Age $< 64.0000$ then Severity = bénigne	Margin $\geq 4.5000$ then Severity = bénigne
Age $\geq 64.0000$ then Severity = maligne	Age $\geq 60.5000$ then Severity = maligne
BIRADS $\geq 2.5000$	Age $\geq 63.5000$
Age $< 68.5000$	Age $< 64.5000$ then Severity = bénigne
Age $< 48.5000$	Age $\geq 64.5000$
Margin $< 4.5000$ then Severity = maligne	Age $< 65.5000$ then Severity = maligne
Margin $\geq 4.5000$ then Severity = maligne	Age $\geq 65.5000$
Age $\geq 48.5000$	Age $< 67.5000$

---

Margin < 4.5000	Age < 44.5000
Age < 66.5000 then Severity = bénigne	Age < 43.0000 then Severity = bénigne
Age >= 66.5000 then Severity = maligne	Age >= 43.0000 then Severity = maligne
Margin >= 4.5000	Age >= 44.5000 then Severity = bénigne
Age < 66.5000 then Severity = maligne	Margin >= 4.5000 then Severity = maligne
Age >= 66.5000 then Severity = maligne	Shape >= 2.5000
Age >= 67.5000	Margin < 1.5000 then Severity = bénigne
Margin < 4.5000 then Severity = maligne	Margin >= 1.5000
Margin >= 4.5000 then Severity = maligne	Age < 59.5000
Age >= 68.5000 then Severity = bénigne	Age < 33.5000
Age >= 69.5000	Age < 32.0000 then Severity = maligne
Age < 74.0000 then Severity = maligne	Age >= 32.0000 then Severity = bénigne
Age >= 74.0000	Age >= 33.5000
BIRADS < 3.5000 then Severity = bénigne	Age < 41.5000 then Severity = maligne
BIRADS >= 3.5000	Age >= 41.5000
Age < 76.5000	Age < 42.5000
Margin < 3.5000 then Severity = maligne	Margin < 4.5000 then Severity = maligne
Margin >= 3.5000 then Severity = bénigne	Margin >= 4.5000 then Severity = bénigne
Age >= 76.5000 then Severity = maligne	Age >= 42.5000
Density >= 3.5000 then Severity = maligne	Age < 46.5000 then Severity = maligne
BIRADS >= 4.5000	Age >= 46.5000
Density < 3.5000	Shape < 3.5000 then Severity = maligne
Density < 1.5000 then Severity = maligne	Shape >= 3.5000
Density >= 1.5000	Margin < 2.5000
Age < 60.5000	Age < 57.5000 then Severity = maligne
BIRADS < 5.5000	Age >= 57.5000 then Severity = bénigne
Shape < 2.5000	Margin >= 2.5000
Margin < 3.5000 then Severity = maligne	Age < 57.5000
Margin >= 3.5000	Density < 2.5000 then Severity = maligne
Age < 41.0000 then Severity = maligne	Density >= 2.5000
Age >= 41.0000	Age < 52.5000
Margin < 4.5000	Age < 49.5000

---

Margin < 4.5000	Age >= 60.5000
Age < 48.5000 then Severity = maligne	BIRADS < 5.5000
Age >= 48.5000 then Severity = maligne	Margin < 3.5000 then Severity = maligne
Margin >= 4.5000 then Severity = maligne	Margin >= 3.5000
Age >= 49.5000 then Severity = maligne	Age < 83.5000
Age >= 52.5000	Age < 72.5000
Age < 56.5000	Age < 70.5000
Age < 55.5000	Age < 66.5000
Margin < 4.5000	Age < 62.5000 then Severity = maligne
Age < 53.5000 then Severity = maligne	Age >= 62.5000
Age >= 53.5000	Shape < 2.0000 then Severity = maligne
Margin < 3.5000	Shape >= 2.0000
Age < 54.5000 then Severity = bénigne	Shape < 3.5000
Age >= 54.5000 then Severity = maligne	Age < 64.0000 then Severity = maligne
Margin >= 3.5000	Age >= 64.0000 then Severity = maligne
Age < 54.5000 then Severity = maligne	Shape >= 3.5000
Age >= 54.5000 then Severity = maligne	Age < 63.5000 then Severity = maligne
Margin >= 4.5000	Age >= 63.5000
Age < 53.5000 then Severity = maligne	Age < 65.5000
Age >= 53.5000	Margin < 4.5000
Age < 54.5000 then Severity = maligne	Age < 64.5000 then Severity = maligne
Age >= 54.5000 then Severity = maligne	Age >= 64.5000 then Severity = maligne
Age >= 55.5000 then Severity = maligne	Margin >= 4.5000 then Severity = maligne
Age >= 56.5000	Age >= 65.5000
Margin < 3.5000 then Severity = maligne	Margin < 4.5000 then Severity = maligne
Margin >= 3.5000 then Severity = bénigne	Margin >= 4.5000 then Severity = maligne
Age >= 57.5000	Age >= 66.5000 then Severity = maligne
Density < 2.5000	Age >= 70.5000
Age < 58.5000 then Severity = maligne	Shape < 3.0000 then Severity = bénigne
Age >= 58.5000 then Severity = bénigne	Shape >= 3.0000
Density >= 2.5000 then Severity = maligne	Margin < 4.5000 then Severity = maligne
Age >= 59.5000 then Severity = maligne	Margin >= 4.5000
BIRADS >= 5.5000 then Severity = maligne	

---

Age < 71.5000 then Severity = maligne  
Age >= 71.5000 then Severity = maligne  
Age >= 72.5000 then Severity = maligne  
Age >= 83.5000  
Age < 84.5000 then Severity = bénigne  
Age >= 84.5000  
Age < 86.5000  
Age < 85.5000 then Severity = maligne  
Age >= 85.5000  
Shape < 3.5000 then Severity = maligne  
Shape >= 3.5000  
Margin < 4.5000 then Severity = bénigne  
Margin >= 4.5000 then Severity = maligne  
Age >= 86.5000 then Severity = maligne  
BIRADS >= 5.5000  
Shape < 3.5000 then Severity = bénigne  
Shape >= 3.5000 then Severity = maligne  
Density >= 3.5000  
Age < 46.5000 then Severity = maligne  
Age >= 46.5000  
Age < 68.5000 then Severity = bénigne  
Age >= 68.5000 then Severity = maligne

---

# Bibliographie

- [1] Gilles Landry. Le cancer, Deuxième édition, Société canadienne du cancer, canada, Pg 6-7, 2008.
- [2] Dominique Huas, Médecin généraliste, Tumeur, [www.docteurlic.com](http://www.docteurlic.com). France, 2010.
- [3] André Morizet. Comprendre le Cancer du sein. L'institut national du cancer, avenue, Pg 9, 2007.
- [4] Professeur X Pivot, Professeur M Marty, Docteur M Espié. Diagnostiquer une tumeur du sein : argumenter l'attitude thérapeutique et justifier le suivi du patient CHU de Besançon, Hôpital Saint Louis-Paris, Pg1-2, 2006.
- [5] Pr Megueni, Dr Chaabni. Une étude portée sur le cancer du sein chez les femmes hospitalisées en gynécologie, CHU Tlemcen, 2006.
- [6] <http://www.chuv.ch>, 21-11-2013 à 16:33.
- [7] Imene Cheikhrouhou, Université d'Evry-Val d'Essonne. Description et classification des masses mammaires pour le diagnostic du cancer du sein, Pg10-26, 2013.
- [8] <http://archive.ics.uci.edu/ml/datasets/Mammographic+Mass>, 18-01-2016 à 11:25.
- [9] Brigitte Séroussi, Jacques Bouaud. Systèmes informatiques d'aide à la décision en médecine: panorama des approches utilisant les données et les connaissances, Sorbonne Universités, UPMC Université Paris 06, Pg2, Janvier 2015.
- [10] Laurent HENRIET, Systèmes d'évaluation et de classification multicritères pour l'aide à la décision, Université Paris Dauphine, Paris, Pg20, 21, 2000.
- [11] Benchaib Yassmine, RDF cours 4 IBM L3 S1 : la classification, Université abou bakr belkaid, Tlemcen 2013.
- [12] <http://archive.ics.uci.edu/ml/datasets/Mammographic+Mass>, 25-01-2016 à 18.45.
- [13] Simone A. Ludwig, Prediction of Breast Cancer Biopsy Outcomes Using a Distributed Genetic Programming Approach, Department of Computer Science, University of Saskatchewan, Canada, Pg 04, 2010.

- 
- [14] Ali KELES, Ayturk KELES, Extracting fuzzy rules for the diagnosis of breast cancer, Department of Computer Education and Instructional Technology, Faculty of Education, Agri Ibrahim C ecen University, Agri, Turkey, 2013.
- [15] Shu-Ting Luo, Bor-Wen Cheng, Diagnosing Breast Masses in Digital Mammography Using Feature Selection and Ensemble Methods, National Yunlin University of Science and Technology, Taiwan, 2010.
- [16] Alaa M. Elsayad, Predicting the Severity of Breast Masses with Ensemble of Bayesian Classifiers, Department of Computers and Systems, Electronics Research Institute, Egypt, 2010.
- [17] Benaki Lairenjam, Yengkhom Satyendra Singh, Hybrid Neural Network for Classifying Mammographic Data, Department of Mathematics, Haramaya University, Dire Dawa, Ethiopia, 2015.
- [18] Benaki Lairenjam, Siri Krishan Wasan, Neural Network with Classification Based on Multiple Association Rule for Classifying Mammographic Data, Department of Mathematics, Jamia Millia Islamia New Delhi, India, 2009.
- [19] Kathleen H. Miao, and George J. Miao, Senior Member, IEEE, Mammographic diagnosis for breast cancer biopsy predictions using neural network classification model and receiver operating characteristic (ROC) curve evaluation, Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Bioinformatics (JBIO), September Edition, 2013.
- [20] Simone A. Ludwig, Prediction of Breast Cancer Biopsy Outcomes Using a Distributed Genetic Programming Approach, Department of Computer Science, University of Saskatchewan, Canada, 2010.
- [21] Guillaume CALAS. Études des principaux algorithmes de data mining, Ecole d'ingénieur en informatique EPITA, France, Pg 01, 2009.
- [22] Cécile Capponi. Arbres de décision, Université Aix-Marseille, Pg 03, Octobre 2012.
- [23] Bertrand LIAUDET. Cours de data mining 5 : modélisation supervisée les arbres de décision, EPF – 4/ 5ème année - Option Ingénierie d’Affaires et de Projets – Finance. France, Pg 02, 09, septembre 2008.
- [24] Sauleau EA. Arbres, Pg 05 – 07, 2007.
- [25] [http://Arbre de décision \(apprentissage\) Wikipédia.html](http://Arbre de décision (apprentissage) Wikipédia.html), 18-02-2016 à 13:56.

- 
- [26] Jérôme Azé. Formation par Apprentissage, Polytech'Paris-Sud, Département informatique, Paris, Pg 09-12, février 2010.
- [27] Shomona Gracia Jacob et R.Geetha Ramani. Evolving efficient clustering and classification patterns in lymphography data through data mining techniques, Department of Computer Science and Engineering, Rajalakshmi Engineering College Department of Information Science and Technology, College of Engineering, Guindy, Anna University, Chennai, Pg 120, 2012.
- [28] [http://:Algorithme ID3 Wikipédia.html](http://AlgorithmesID3Wikipedia.html), 04-03-2016 à 12:25.
- [29] Hannachi Sabrine, Rabia Maroia, Projet fouille de données, Université Claude Bernard, Lyon, Pg 13, 2014.
- [30] Rimah Amami, Dorra Ben Ayed, Nouredine Ellouze, Application de la Méthode Adaboost à la Reconnaissance Automatique de la Parole, Département de Génie Electrique, ENIT, Tunis, Tunisie Pg 04, 2011.
- [31] [http:// docs.rapidminer.com /studio /operators /modeling /classification and regression/tree induction/random tree.html](http://docs.rapidminer.com/studio/operators/modeling/classification_and_regression/tree_induction/random_tree.html), 04-04-2016 à 09:40.
- [32] Asma HAMMYANI, Soumia ALLIOUA, Amélioration des forêts aléatoires : Application au diagnostic médical, Univ. Tlemcen, Pg 09, 2013.
- [33] Mélanie Glasson-Cicognani & André Berchtold, Imputation des données manquantes: Comparaison de différentes approches, Université de Lausanne, Institut de Mathématiques Appliquées SSP, Anthropole, CH - 1015 Lausanne, Pg 02, 2010.
- [34] Francisco Herrera, Data Mining and Soft Computing, Research Group on Soft Computing and Information Intelligent Systems (SCIIS) Dept. of Computer Science and A.I. University of Granada, Spain, Novembre, Pg 191, 2008.